

Supporting good citizenship?

With dialogue between science and society all the rage, it would seem sensible to start young. A scheme for school education means well, but will fizzle out without the required commitment from government.

On the face of it, the new move to include science in the teaching of citizenship in English schools from September this year appears promising. What better way to address the issues of the impact of science on society, and the ability of society to engage better in 'scientific' debate, than to catch your citizens young?

But this approach has its weaknesses (see page 3), which threaten its chances of hitting the target. First, and probably crucially, the new focus on citizenship — into which science seems to have been squeezed as an afterthought — comes with no additional resources of time or money for teachers. Educational researchers, head teachers and pupils will draw their own conclusions about the real importance of citizenship education in the government's eyes.

Second, there are questions over how well specialist teachers of science will be able to initiate and manage classroom discussion with little or no formal training in these skills. Good teachers already introduce discussion into their lessons — but it will strike many science teachers in England as ironic that, a decade or so into the restrictive autocracy of the national curriculum, they are now expected to be more flexible and free-thinking.

Many students are drawn to science by its logic and the sense that a definitive approach can be found to complicated questions. But school-children also relish an argument. Indeed, they are notorious activists on

some issues — the environment was discussed in classrooms long before mainstream politicians and oil companies discovered it. We should not underestimate the eagerness and ability of pupils to address seemingly remote and complex social issues while learning the hard scientific facts about cloning, food production and genetic testing.

Much of the citizenship agenda seems intended to address a growing reluctance among young people in Britain to participate in politics. Politicians like to label this as apathy, but many non-voters argue that they are making a deliberate choice not to take part in what they see as a remote process with little bearing on their everyday lives. Including the ethics of science in formal attempts to redress this could risk alienation by association. One science teacher with 20 years' experience said: "I've been including discussions like this in lessons for years, but I'd never tell the kids it was citizenship."

Those behind the new initiative stress its "light touch" and its "flexibility", meaning that it can be jemmied into existing lessons when teachers find the opportunity or time. Inevitably, some will try harder than others. But a way of introducing lively and relevant discussion into all science lessons already exists — and not as an afterthought. Science teachers should be given the training and the time to include it as a matter of course. Educationalists should listen to teachers and loosen the shackles of the national curriculum. ■

How to encourage the right behaviour

The sharing of published data and research materials is a requirement of many journals. Can we ensure that it happens?

"When I started on this I thought it would be boring. Having seen such distinguished people work hard at it for a day, I see it isn't." So said Tom Cech, president of the Howard Hughes Medical Institute, at a gathering last week of academics, industrialists, administrators and editors at the US National Academy of Sciences seeking a consensus on the role of journals in the availability of published data and research materials. Cech's degree of interest is relevant because he heads an academy panel that is due to report on the issues of availability within a few months.

Virtually the only consensus to be found was of the motherhood-and-apple-pie variety. The stand-off persists between those seeking free access to all published data and those, not only in the database industry, who want journals to allow published data to be deposited in subscriber-only databases. As somebody remarked, not only the Devil but God is in the details: if you want to set out principles of access, you have to get into the context — a simple statement of principle cannot do justice to the varied state of the experimental arts in different fields and the differing expectations of the various research communities.

Equally important is the largely hidden problem of non-compliance with journals' guidelines on data sharing. Many refusals to share materials probably don't get reported, and where there are complaints, they go variously to funding agencies, authors' employers and journals, so that the full extent of non-compliance is unknown. Should there be a standard set of procedures for complainants to follow?

This journal's experience matches that of several editors at the meeting: a letter from the journal to an author is generally enough to overcome a reluctance to share, except where there are genuine obstacles to availability of certain types of material. A journal's editor is thus a sensible first stop for a researcher facing an author's stubborn refusal to share according to the journal's guidelines.

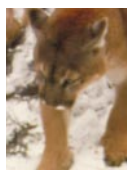
But there can be cases where a quiet word from the editor is inadequate. Whom do you complain to then? How does a dissatisfied researcher or editor take the case further? A threat from a funding agency (usually identified in the paper) probably carries the greatest clout. But can universities or other employers of recalcitrant researchers, eager to protect their laboratory's competitiveness, be relied on to follow the rules?

Cech also pointed out, quite rightly, that although many journals issue strong statements of principle about sharing, they give little or no guidance about how to put pressure on authors who refuse to comply. Point taken. We have now added to our website the simple instruction that anybody who fails to obtain data or materials in accordance with our Guide to Authors should contact the Editor at nature@nature.com, who will ensure that action is taken. That, at least, is a start. But for these and more serious cases of misconduct, funding agencies and employers should be more transparent about whom people should contact in cases of persistent or extreme bad behaviour, and their commitment to act on complaints. ■





Mitch's metrics
The White House reveals assessment criteria for science
p4



Lion watching
Predation studies get go-ahead despite concerns
p5



AIDS vaccine
HIV trial set to proceed as rival is abandoned
p6



Badger blight
Genome of bovine tuberculosis bacterium sequenced
p8

Britain banks on embryonic stem cells to gain competitive edge

David Adam, London

Britain is staking an aggressive claim for global leadership in stem-cell research by offering all of its biologists easy access to newly derived, high-quality embryonic stem-cell lines.

On 27 February, the Medical Research Council (MRC) confirmed its plans to establish a public cell bank to characterize and store adult and embryonic stem cells. It will also distribute these cells for a nominal charge to both academic and commercial research groups.

Coming on top of recent British regulations, which are among the first in the world to endorse publicly funded research on embryonic stem cells, science administrators hope that the cell bank will help to foster world leadership in the emerging field of regenerative biology.



Wise investment? A new cell bank will make stem cells far more readily available to researchers.

"I think it's a very smart thing to do, it's a very positive move," says Ron McKay, a leading stem-cell researcher at the US National

Institute of Neurological Disorders and Stroke. "The position of Britain now in this field is critical."

The bank, which should open in 2003, will house embryonic stem-cell lines that are expected to be derived in Britain over the coming months, together with some imported lines already existing. The MRC announced its plans to build the bank just as a House of Lords select committee on stem-cell research published a report recommending that such a facility be established.

No UK group has yet derived embryonic stem cells, although supporters argue that Britain now has a regulatory framework that will give its researchers an inside track to the early clinical application of such cells. The House of Lords report also clears the way for 'therapeutic' cloning of a patient's cells to create early-stage embryos as a source of stem cells, which Britain permits under a law passed at the end of 2000.

Stem cells taken from early embryos can turn into any type of tissue, and researchers believe the cells could be used to treat a range of conditions including Parkinson's disease, hepatitis, diabetes and leukaemia.

"The MRC supports this area of research and believes that it has real potential," says George Radda, the council's chief executive. "The stem-cell bank will allow researchers to explore this enormous potential."

Radda says that researchers using MRC funds to derive embryonic stem cells will be

Call for cloning ban splits UN

Erika Check, Washington

The United States and its European allies are once again on a collision course over an international agreement. This time, the bone of contention is a proposed global ban on human cloning, under consideration at the United Nations (UN).

The UN committee began deliberating the proposal last week. But on the first day of the talks, the American delegation said that any anti-cloning agreement should also ban the 'therapeutic' cloning to form human embryos for research purposes.

China, Japan and some European nations already permit such research, and their representatives will argue that the US position would prevent scientists from carrying out potentially beneficial work.

The UN set up the Ad Hoc Committee on an International Convention Against the Reproductive Cloning of Human Beings in December in response to a request made last August by France and Germany for a ban on human reproductive cloning.

After the meeting, committee chair Peter

Tomka of Slovakia said that the group did not reach a consensus and would meet again in September. Shortly after that meeting, the UN's General Assembly is expected to decide whether to pass a resolution against cloning.

The US delegation's declaration, which states that "to ban 'reproductive' cloning effectively, all human cloning must be banned", reflects the views of the administration of President George Bush. But the United States has no laws against either reproductive or therapeutic cloning.

Observers say that the US position is unlikely to win the support of the UN. But it could serve to deadlock the process and prevent the reproductive cloning ban from going ahead.

"This could be a Kyoto of the embryos," says Arthur Caplan, director of the Center for Bioethics at the University of Pennsylvania, who advised the Clinton administration on bioethics and is working with the UN ad hoc committee. The United States "is out of step with the world's position", he says.

required to place the resulting cell lines in the new bank; other UK funding agencies are expected to introduce similar guidelines. Providing the cells meet strict quality-control criteria, they will then be frozen, stored and made available to anyone who wants them for projects approved by an advisory committee.

At first, only groups in Britain will be eligible to apply, but Radda says that the MRC "in principle" would be willing to make them available to overseas groups.

The bank's location has yet to be determined, although Radda says it will not be at a university, a commercial organization or an MRC research institute, as the MRC wants to avoid a situation in which researchers involved in stem-cell research have a say in decisions about who should be supplied with the cells. A decision on location will be made in July, the MRC says.

The ease with which UK researchers will be able to obtain identical cell lines from the bank will help different groups to compare their results, says Richard Gardner, an embryologist at the University of Oxford and chairman of the Royal Society working group on stem-cell research.

Dozens of embryonic stem-cell lines have already been derived worldwide, but Gardner says their quality depends on the conditions under which they were derived — which were highly variable. To provide the best research resource, he suggests, "one essentially needs to start again under very carefully defined conditions" ■

White House sets three-point performance plan for science

Geoff Brumfiel, Washington

The White House Office of Management and Budget (OMB) has unveiled its long-awaited criteria for evaluating basic research supported by the US government.

Research organizations had worried that the criteria would force agencies to quantify the output of basic research programmes and so would damage high-risk projects (see *Nature* **413**, 5; 2001).

At a meeting at the National Academy of Sciences in Washington on 27 February, OMB officials said that basic research will be measured by three criteria: the quality of the research, its relevance to the funding agency, and its performance based on defined goals and measures.

They added that research programmes should be reviewed every three to five years, and that each agency should set appropriate standards against which its overall research performance can be assessed annually. These measures have yet to be set, and it remains unclear how they will influence agencies' future budgets.

Mitch Daniels, director of the OMB, told the meeting that he was strongly committed to the criteria for basic science. "Even in this most speculative area of government investment, our decisions cannot be immune from standards and quality," he said.



Setting standards: Mitch Daniels is a keen advocate of performance criteria for research.

The meeting did manage to provide some assurance that the process will be managed carefully. "The nervousness people had was that this was some kind of ideological juggernaut," says David Goldston, chief of staff for the science committee in the House of Representatives. "One of the things the OMB did at the meeting was to put that to rest."

But concerns remain. Mildred Dresselhaus, a professor of electrical engineering at the Massachusetts Institute of Technology, warned the meeting that the OMB's calls for regular reports might hurt areas of research that flounder for years before succeeding. "How will we incorporate failure into the criteria?" she asked. ■

♦ www7.nationalacademies.org/gpra/index.html

Stem-cell reverse angers Australian biologists

Carina Dennis, Sydney

Hints that the Australian government is planning a national ban on the generation of new human embryonic stem-cell lines have drawn fire from biologists and their backers in some of the nation's state governments.

Kevin Andrews, the minister for ageing — who is responsible for stem-cell research and

cloning policy — has given the cabinet a set of proposals which, according to reliable reports, include a ban on the use of human embryos to extract new stem-cell lines.

Andrews has refused to comment on his submission, but has made his opposition to embryo research clear in the past.

Although the government has not finalized its position, supporters of the research are angered by what they see as an attempt to overturn the findings of a two-year parliamentary inquiry into cloning and stem-cell research, completed last September. The inquiry committee — which Andrews chaired — was unanimous on all issues except the use of discarded embryos in stem-cell research, which was approved by a majority. Andrews was one of the minority opposed to such work.

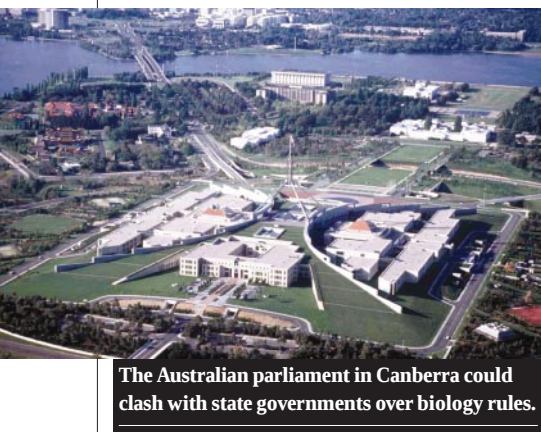
Martin Pera, of the Monash Institute of Reproduction and Development, Melbourne, whose group was one of the first to isolate human embryonic stem-cell lines, argues that additional lines are crucial to offer more

options to researchers, and to ensure that genetically diverse lines are available for clinical application.

Robert Jansen, a professor at the University of Sydney and head of Sydney IVE, an *in-vitro* fertilization clinic, says that restrictive laws in stem-cell research "will put Australia into a backwater" and force researchers abroad.

Under the Australian constitution, health-related matters normally fall under the jurisdiction of the states, not the federal government. The questions of whether federal or state governments will legislate, and of which stem-cell research and cloning technologies should be permitted, will be discussed at the next Council of Australian Governments' meeting in April.

The federal government and the states — several of which are governed by the opposition Labor Party — look set to clash on the issue, probably delaying implementation of the agreement reached last year to establish nationally consistent rules by June. ■



The Australian parliament in Canberra could clash with state governments over biology rules.

Protests fail to block mountain-lion surveys

Rex Dalton, San Diego

Extensive studies are under way to investigate the impact of predation by mountain lions on wildlife in the western United States. The results are expected to strongly influence the management of scarce and endangered species in the region.

Having defeated a legal challenge by environmentalists in the federal court in Portland last month, biologists with the Oregon Department of Fish and Wildlife are proceeding with a \$5.2-million study of two elk populations, whose numbers have been declining in recent years.

And wildlife biologists with the Arizona Game and Fish Department have just received final approval from the federal government for fieldwork on the effect of lions on a desert bighorn sheep population northeast of Phoenix. The sheep population has recently plummeted, after a successful restoration effort begun 20 years ago.

In both cases, state biologists may kill some lions during the studies to reduce the lion population and determine whether this improves elk or sheep survival. But environmentalists have sought to block both studies because of the proposed killings.

The studies are expected to answer pressing questions about whether it is lions, environmental change or disease that lies behind declining wildlife populations.

The results could have implications for the conservation of many species. Mountain lions, which weigh between 50 and 85 kilograms and are better known as cougars or pumas, head the food chain in many parts of North America. In recent decades, their populations have blossomed, presenting a major complicating factor for biologists seeking to save or restore wildlife.

In California, for instance, the last remaining population of around 250 Sierra Nevada bighorn sheep — which are on both state and federal endangered-species lists — has been seriously threatened by predation by lions in the mountains near Bishop, south of Lake Tahoe.

The studies in Oregon and Arizona will be closely watched by biologists interested in the Californian sheep. "We could never do these studies in California," says biologist Vern Bleich of the California Department of Fish and Game, citing political resistance to work that might lead to calls for limits on the lion population. California outlaws lion hunting, which is legal in Oregon and Arizona and will be factored into the research.

Environmentalists' objections to the Oregon study peaked last month when the Sierra Club and other groups sued the US Fish and Wildlife Service, alleging that there had been inadequate environmental



Killer in the sights? Environmentalists want studies halted because of plans to cull the lions.

assessment of the study. But the court in Portland ruled on 19 February that it would not halt the research.

Under the five-year study, biologists will fit some of the lions with radio collars and monitor their movements, together with those of Rocky Mountain elk in national

forests in the northeast part of the state and Roosevelt elk in the southwest.

In the Arizona project, biologists will decide in the autumn whether they need to kill any lions. A total of 12 lions are to be eliminated in the study area, but nine have already been killed by sports hunters. ■

Citizenship gets a science angle

David Adam, London

English schoolchildren will soon be grappling with the social issues surrounding topics such as the growing of genetically modified food under government plans to give 'citizenship' lessons a formal place in the school timetable.

But with no extra resources and little training for science teachers in such classroom debates, some science education experts are questioning whether the scheme will work.

From September, English schools will be required to incorporate discussions on social

issues into classes. Science education experts are keen that science teachers hold discussions on topics such as genetically modified crops and nuclear power. "Citizenship education is a golden opportunity to get young people to think about science from a more informed perspective," says Peter Finegold of the Wellcome Trust, Britain's largest medical research charity.

On 28 February, the trust, together with Britain's Association for Science Education, held a conference in London to promote the scientific aspects of the plan. Most teachers at the meeting were enthusiastic, despite worries over how the new subject would be incorporated.

But some education researchers suggested that science teachers could be less suited to promoting such discussions than colleagues who teach other subjects. Last year, a Wellcome Trust report found that most classroom discussions of biomedical issues take place in humanities lessons.

"In general, science teachers feel that they lack the skills, confidence and time to initiate classroom discussion," says Ralph Levinson of the University of London, co-author of the Wellcome report. "And the scientific facts appear to be incidental to the teaching of these issue-based topics in the humanities." ■



Army HIV vaccine to undergo clinical trial as rival is halted

AP PHOTO SAKCHAI LALIT



Testing times: Thai volunteers will soon be trying out an HIV vaccine developed by the US army.

Erika Check, Washington

The National Institutes of Health (NIH) has decided not to proceed with a full-scale clinical trial of its most advanced candidate HIV vaccine because of the vaccine's poor showing in smaller safety and efficacy trials.

But ironically, the NIH will now find itself responsible for running a large-scale study — known as a phase III trial — of a very similar HIV vaccine. This was developed by US Army researchers at the Walter Reed Army Institute of Research in Washington DC, in cooperation with the Thai government.

A January directive from the White House Office of Management and Budget instructs the NIH's National Institute of Allergy and Infectious Diseases (NIAID) to assume responsibility for the army's trial, along with all military HIV/AIDS research, in October. That trial will involve about 16,000 young Thai volunteers and is set to start at the end of the year. The NIH will provide funding for the trial, but other details of the transfer have still to be agreed by the NIH and the army.

Critics have complained that the dual civilian and military trials were a waste of resources (see *Nature* 415, 365–366; 2002). Both use a 'prime-boost' strategy, which delivers HIV viral proteins in two different ways. The vaccinees are first given a shot of canarypox virus containing HIV genes that code for proteins from the virus' outer coat. This 'prime' shot is followed up by a 'boost' shot of engineered pieces of the protein gp120, which is also found in HIV's coat.

Supporters of the NIH vaccine effort were disappointed that their vaccine had come up short in the smaller, phase II trial. "This is really very unfortunate," says David Baltimore, a molecular biologist and president of the California Institute of Technology, who

chairs the NIH's AIDS vaccine research committee. "The development of the HIV vaccine has been excruciatingly slow. It has turned out to be so much more difficult than we originally imagined."

The NIH phase II trial aimed to find out what level of immune response protected vaccinees from the virus. To do that, the vaccine needed to produce a particular immune response in at least 30% of the volunteers. But the civilian phase II trial "missed that mark by a fair amount", says Anthony Fauci, director of the NIAID. Because the low response will not allow investigators to find out what creates immunity to the virus, Fauci says, that trial did not meet its goal and will not continue.

In contrast, the army's phase III trial will simply evaluate whether the vaccine protects against HIV infection. A spokesperson at Walter Reed says that the appropriate immune response was provoked in 25% of the volunteers in the phase II military trial. But this result is not strictly comparable to the civilian trial — the army trial used different methods to measure immune responses, for example, and took more frequent measurements in volunteers.

Some researchers think that the army vaccine will be no more effective than the civilian vaccine. But there are other reasons why the army's phase III trial will continue, they say, including the fact that it builds on years of cooperation between the American and Thai governments.

"This trial is more than just a vaccine trial," says Beatrice Hahn, an HIV researcher at the University of Alabama at Birmingham. "They have made such an investment in training, infrastructure, technology transfer, assays and equipment that at this point it's impossible to pull the plug, and no one would want to do so."

Congress seeks to keep Sea Grant in its current harbour

Virginia Gewin, Washington

Congress is poised to spike a White House plan to transfer the Sea Grant programme from the National Oceanic and Atmospheric Administration (NOAA) to the National Science Foundation (NSF).

The National Sea Grant College Program, which supports research into pollution, sea-food safety and fisheries management, is strongly supported by members of Congress from coastal regions, who want NOAA to stay in charge.

At a hearing on 28 February of a subcommittee of the House of Representatives' science committee, scant support was voiced for the transfer, and four of the five witnesses strongly opposed it. They argued that the hallmark of Sea Grant — its links to local communities and organizations — would be lost if it moved to the NSF.

The House resources committee has already approved a bill that would keep the programme at NOAA for another five years, and the science committee is likely to follow suit. Rick DeVoe, president of the Sea Grant Association, says a bill in the Senate will be developed soon.

"We are optimistic that the House will be able to bring legislation to the floor in the coming months that will reauthorize Sea Grant at NOAA and provide the programme with a robust funding increase," says Carolyn Thoroughgood, president of the Consortium for Oceanographic Research and Education, which lobbies for oceanographic research. But the funding increase, at least, remains uncertain, congressmen say.

If the transfer is derailed in Congress, it will be the second time this year that a plan by the White House Office of Management and Budget to widen the NSF's remit has failed. Its proposed transfer to the NSF of three Smithsonian Institution research centres was abandoned in January (see *Nature* 415, 252; 2002).

AP PHOTO BRIAN BRANCH-PRICE



Pollution in Maryland's Chesapeake Bay is monitored by the Sea Grant programme.

Bubble fusion dispute reaches boiling point

Geoff Brumfiel, Washington

Claims that nuclear fusion has been achieved in a 'table-top' experiment have become mired in controversy before the findings have even appeared in print.

The authors of the paper believe their experiment, reported in this week's issue of *Science*, is a breakthrough that might one day help to fulfil the elusive dream of fusion power. But opponents dispute the findings and say that the paper carries echoes of the 'cold fusion' fiasco of the late 1980s, in which two researchers claimed to have achieved nuclear fusion using an electrolytic cell.

The authors of the paper, led by Rusi Taleyarkhan of the Oak Ridge National Laboratory in Tennessee, claim to have achieved fusion inside a vibrating beaker filled with acetone, in which the hydrogen atoms were replaced with deuterium. This is a heavier isotope of hydrogen containing a neutron, in addition to a proton, in its nucleus.

The researchers bombarded the sample with neutrons to 'seed' the formation of tiny bubbles. When they shocked the beaker with waves of sound, bubbles formed, expanded and suddenly collapsed, releasing flashes of light. This overall phenomenon, known as sonoluminescence, is thought to cause a momentary surge in temperature inside the bubbles that some physicists have argued might be sufficient to sustain nuclear fusion reactions.

This is exactly what Taleyarkhan and his colleagues claim to have recorded. They argue that — under the right experimental conditions — the temperature in the bubbles rose to more than 1 million kelvin, allowing deuterium nuclei to undergo fusion reactions. As evidence, the team cites the production of tritium — a form of hydrogen containing two neutrons — and high-energy neutrons (R. P. Taleyarkhan *et al. Science* **295**, 1868–1873; 2002).

But not everyone is convinced, and some of the loudest critics are also at Oak Ridge. As part of an internal review, nuclear physicists Dan Shapira and Michael Saltmarsh replicated the experiment last July while the paper was being reviewed for publication. "We substituted our own detector and a much more sophisticated data-acquisition system to see if we could see what they saw," says Saltmarsh, "and we didn't."

Saltmarsh and Shapira claim that the neutron production is an order of magnitude too small to be consistent with the tritium production reported by Taleyarkhan. They also argue that the paper does not prove that the neutron emissions coincide with the sonoluminescent flashes.

After completing reports to the original authors and Oak Ridge management, Saltmarsh and Shapira put the matter out of



Bubble trouble: a team of physicists claims to have achieved nuclear fusion inside small bubbles created with apparatus similar to that shown above, but the results are facing strong challenges.

their minds. "We thought this was all put to bed," Saltmarsh says.

But three weeks ago, they found out that Taleyarkhan's paper was going to appear in *Science*. Saltmarsh and his colleagues scrambled to publish their results, but their data had not been peer reviewed. "It would have been nice to get our results published at the same time, but we had no warning," Saltmarsh says.

Heated debate

Nonetheless, their results, posted on an Oak Ridge website, are cited in Taleyarkhan's paper together with a similar citation to the fusion team's rebuttal, which is also posted on the Internet. These non-peer-reviewed articles are referred to in the final paper at the insistence of *Science*'s editorial team, following representations from senior managers at Oak Ridge.

Other experts are also sceptical of the fusion experiment's results. "I've reviewed this paper from its first version, and I unfortunately concluded that the experimental technique is flawed," says Seth Putterman of the University of California, Los Angeles.

Putterman asserts that the beam of neutrons used by Taleyarkhan's team to seed the bubbles would be indistinguishable from any neutrons produced in a fusion reaction. "The worst way to look for neutrons is to flood your room with them," he says.

He also has serious concerns about the tritium measurements, arguing that deuterated acetone would naturally contain some tritium, which could become concentrated through processes that have nothing to do

with fusion. Even in advance of *Science*'s early release of the paper on 4 March, Putterman's concerns were being echoed by other leading physicists.

But the authors fiercely defend their paper. Richard Lahey of the Rensselaer Polytechnic Institute in New York state, a co-author on the paper, dismisses Putterman's comments about neutron detection. "We spent hours worrying about that," he says, adding that he believes the neutrons produced by fusion would have a higher energy than those used to seed the bubbles.

Lahey also says that the team accounted for any background levels of tritium present in their sample, and detected a clear excess of the isotope during the experiment.

As the debate rages on, Oak Ridge officials have found an uncomfortable seat on the sidelines. "I think it's an exciting experiment," says Lee Riedinger, deputy director of Oak Ridge. "But to be honest, I need to see another measurement before I can decide if the conditions are there for fusion."

Riedinger notes that interest generated by Taleyarkhan's 'table-top' fusion is reminiscent of the furore that surrounded the 1989 'cold fusion' experiment by Stanley Pons and Martin Fleischmann of the University of Utah.

But he points to two important differences: the current paper has undergone extensive peer review, and most physicists agree bubble fusion is, at least in principle, possible. He hopes that the Saltmarsh paper will soon be peer-reviewed and published. ■

► www.ornl.gov/slsite

► www.rpi.edu/~lahey/r/SciencePaper.pdf

Environmentalists' bugbear to head research institute

Copenhagen Bjorn Lomborg, the author of the controversial book *The Skeptical Environmentalist*, has been appointed director of a new institute for environmental research in his native Denmark. His appointment as head of the Institute of Environmental Evaluation was announced on 26 February by the new Liberal-Conservative coalition government.

In his 2001 book, Lomborg uses statistics to challenge widely held beliefs about the state of the global environment. Many environmental scientists strongly disagreed with his conclusions (see *Nature* **414**, 149–150; 2001).

Lomborg has no formal qualifications in environmental science, and scientists and environmental groups in Denmark have criticized the appointment, saying that it reflects the new government's unhelpful approach to environmental issues. As a further example, they cite the decision to lift a nation-wide ban on disposable packaging, which, they say, encouraged recycling.

Bomb tests killed 11,000 in United States, study claims

Washington Almost every person in the United States has been exposed to radiation from nuclear testing, according to a preliminary report by the Centers for Disease Control and Prevention (CDC).

The study estimates the radiation exposure and potential health effects of tests conducted in the United States and elsewhere between 1951 and 1962. Between 10,000 and 200,000 incidents of thyroid cancer and some 11,000 deaths can be linked to the testing, the authors say. "Any person living in the contiguous United States since 1951 has been exposed to radioactive fallout, and all organs and tissues of the body have received some radiation exposure," the study concludes. But some scientists have criticized the methods used to predict the numbers of

deaths resulting from fallout.

The CDC findings are scheduled to be peer reviewed by the National Academy of Sciences, but were made public by Tom Harkin, Democrat senator for Iowa, and posted on a website run by the Institute for Energy and Environmental Research, an environmental pressure group based in Maryland. The academy says it will review the findings and release its own report in six to nine months' time.

♦ www.ieer.org

FDA gains a deputy as no one agrees on head

Washington President George W. Bush seems to have found a way around the deadlock that has left the US Food and Drug Administration (FDA) leaderless for more than a year.

Lester Crawford, a veterinary surgeon and food-safety specialist from the Center for Food and Nutrition Policy in northern Virginia, was made deputy commissioner of the FDA on 25 February. This post does not require congressional approval, unlike the FDA commissioner's job, which remains vacant because of political differences.

The Bush administration has come under fire for failing to fill several top scientific positions at important agencies (see *Nature* **415**, 946; 2002). Crawford is reported to have previously been approached by the Bush administration for the post of commissioner, and could now run the agency indefinitely.

TB gene sequence could save cows and badgers

Paris Humans, cows and badgers could all benefit from the latest genome to be fully sequenced: *Mycobacterium bovis*, the bacterium that causes tuberculosis (TB) in a wide range of wild and domestic animals.

The sequence is almost identical to that of *M. tuberculosis*, the bacterium that causes TB in humans (see *Nature* **393**, 537; 1998). The result means that groups working on vaccines for the different strains could collaborate in the future.



Badger benefits: a TB vaccine could prevent the slaughter of thousands of the animals.

A vaccine for the animal form of TB is badly needed, as the disease costs farmers around the world US\$3 billion a year, say the researchers who carried out the sequencing.

It could also save the lives of thousands of British badgers. The animals are suspected of spreading the disease among cattle, and tens of thousands will be killed over the next few years as part of an experiment into controlling the disease.

The sequence was produced by researchers at the UK government's Veterinary Laboratories Agency, the Sanger Centre near Cambridge, and the Pasteur Institute in France.

♦ www.sanger.ac.uk

US import controls 'can't keep BSE out'

Washington Current controls in the United States are not sufficient to prevent animal products infected with bovine spongiform encephalopathy (BSE) from entering the country, a report by the US General Accounting Office (GAO) has concluded.

BSE was spread among British cattle when parts of infected animals were used in feed for other animals. The US Department of Agriculture began to block imports of animal products from infected countries in 1989, but the report, issued on 26 February, says that there is insufficient manpower to police this ban. Infected feed may also have entered the country before the ban, and the authors question the methods used by the Food and Drug Administration (FDA) to combat this risk. The GAO says that the FDA is not inspecting farms to make sure that they are complying with regulations over use of animal feed.

New cases of BSE are still being reported in Britain, despite a cull of potentially infected animals and a ban on the type of feed that helped to spread the disease. John Wilesmith, an epidemiologist with the government's Veterinary Laboratories Agency, has speculated that imports of meat from infected countries could be the cause of recent new BSE infections. This may explain why Britain has failed to eradicate BSE, Wilesmith said last week.

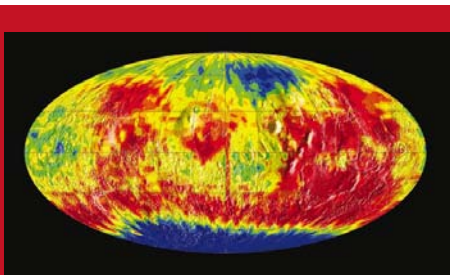
♦ www.gao.gov/new.items/d02183.pdf

Red planet turns blue

Paris NASA's Mars Odyssey spacecraft beamed back its first pictures last week, revealing new panoramas of the red planet and prompting further claims that it contains significant amounts of water.

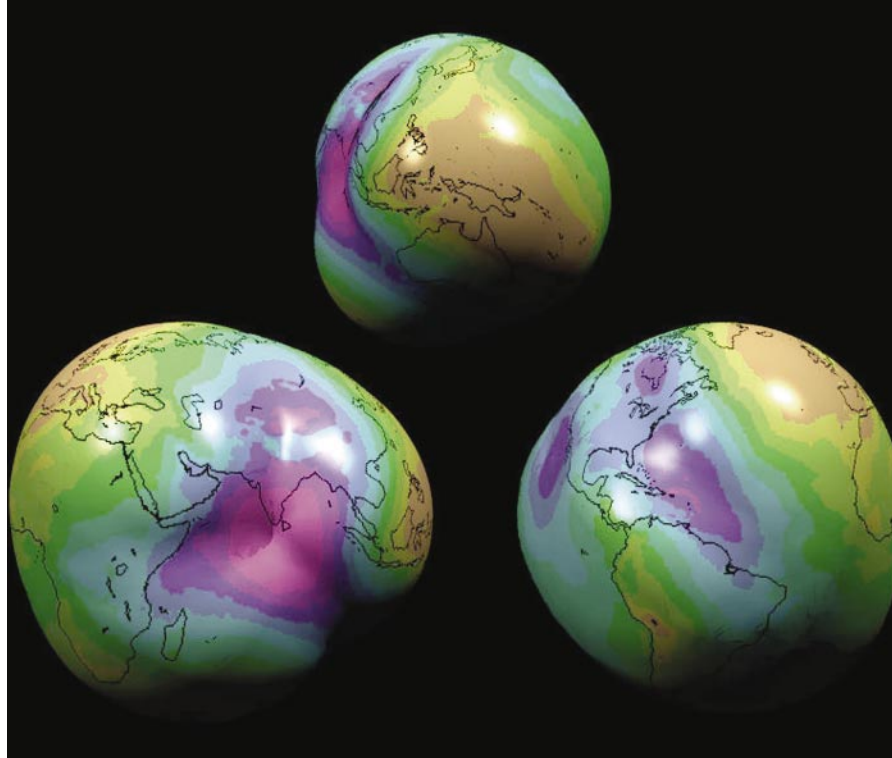
This image shows the intensity of mid-energy neutrons emitted by the planet. The particles are created by collisions between incoming cosmic rays and the planet's surface. Hydrogen near the surface slows down these neutrons, and the blue areas on the image indicate the presence of this element. NASA says the hydrogen is likely to be locked up in ice.

♦ mars.jpl.nasa.gov/odyssey



Amazing grace

Two satellites that will chase one another around the globe are poised to map the Earth's gravitational field with unprecedented accuracy. Researchers are feeling the pull of the data, says David Adam.



Shaping up: this highly exaggerated 'geoid' shows how the pull of gravity varies across the globe.

In an uncertain world, at least the force of gravity seems reliable: apples drop, feathers fall and what goes up must come down. But things are not as clear cut as they seem. Our lumpy Earth is not a perfect sphere, which means that the pull of gravity is stronger in some places than in others — you weigh just a little more in the Rocky Mountains than on a ship in the middle of the Indian Ocean. Nor does gravity keep still. Ocean currents, the flow of magma beneath the Earth's crust and the movement of glaciers all redistribute the planet's mass enough to keep the gravity map in flux.

Earth scientists and oceanographers know roughly what the Earth's gravity map looks like. But until now they have only been able to dream about a map accurate and dynamic enough to monitor phenomena such as ocean currents or the melting of polar ice. This dream is about to come true. On 16 March, an international team of scientists aims to launch the Gravity Recovery and Climate Experiment (GRACE) from Russia's Plesetsk rocket facility, 600 kilometres northeast of St Petersburg. Its five-year mission: to produce the most accurate map of Earth's gravity ever made and to reveal how it changes over time.

GRACE's two trapezoidal satellites will follow identical orbits, one some 220 kilometres ahead of the other. Subtle changes in gravity will affect the leading satellite first, pulling it slightly closer to the Earth, for example, and so fractionally away from its trailing partner. On-board microwave range finders will monitor the distance between the two satellites with extreme precision, reporting back to ground stations up to five times a day. Every 30 days, the satellites will have covered enough of the Earth's surface to

construct a complete planetary gravity map.

"Scientists have been talking about and planning a mission like this for a decade or more," says Philip Moore of the University of Newcastle upon Tyne, UK, who specializes in satellite measurements of gravity. "We are fully expecting the results to be quite startling. It could revolutionize many sciences."

Weighty issues

The US\$150-million mission, a joint project between NASA and the DLR, Germany's aerospace research agency, will reveal new information about the Earth's geology and hydrology. "We'll be able to see various phenomena that involve transporting mass around and how much mass they're actually moving," says Michael Watkins, project scientist for the mission at NASA's Jet

Propulsion Laboratory in Pasadena, California. "These are things that aren't easy to see with any other type of measurement."

Very accurate measurements of gravity in specific places have already been made using gravimeters, sensitive instruments that sense how hard the ground pushes back at them as they are tugged down by gravity. Researchers have produced large-scale maps by combining these ground-based readings with others obtained by tracking the orbital motion of satellites from the ground. In 1999, Germany launched a satellite called CHAMP, which carried Global Positioning System (GPS) equipment to track its position more accurately.

But Watkins says that data from the twin GRACE satellites, which will also carry GPS equipment, will produce a map 100 times more accurate than the best previous effort. The data will be released through free public-access websites at NASA and in Germany after six to nine months, says Christoph Reigber of the GFZ, Germany's national Earth-sciences research centre in Potsdam, who heads the German GRACE team.

One way to think of the gravitational variations that GRACE will measure is a representation known as the geoid. This is an imaginary topographical surface where gravity is constant at every point. Sea water piles up over regions of higher gravity, and the geoid is the surface to which the oceans would naturally settle in the absence of wind, tides and currents. A swimmer in the ocean off the southern tip of India, for example, is some 200 metres lower, or closer to the centre of the Earth, than another swimmer near Borneo.

Climate modellers would love to remove the geoid effect from satellite measurements of sea level and so get at the effects of ocean currents and temperature changes. Both are



Newton's heir: Michael Watkins is set to measure the force that brings apples down to Earth.

more relevant to climate change than the geoid, Watkins says, and a better quantification of the contribution of gravity to these sea-level measurements will allow researchers to eliminate it. "You have to subtract off the geoid to get to that oceanographic part," he says.

Oceanographers such as Philip Woodworth of the Proudman Oceanographic Laboratory near Liverpool, UK, see GRACE as a way to get at deep ocean currents. These currents are part of the global conveyor belt that transports water around the planet, but the ocean depths are inaccessible to standard instruments. Because gravity determines the total mass of water over a particular point on the sea bed, says Woodworth, "it can be used as a surrogate for bottom pressures, which are very important in determining bottom currents". With improved pressure measurements, Woodworth hopes to develop better models of ocean circulation.

Moving pictures

GRACE could help to improve theories about the fate of the Gulf Stream, which brings warm water across the Atlantic to northwest Europe. Some modellers have suggested that melting Arctic ice could turn off the Gulf Stream and plunge northern Europe into a new ice age. Any weakening of the Gulf Stream over GRACE's five-year lifespan should be evident in its sequence of gravity maps of the North Atlantic.

Indeed, the fact that GRACE's maps will be updated every 30 days is a major advance. "Previously gravity has always been treated as a static quantity, and the attitude has been that we measure it one time and then we're done," says GRACE principal investigator Byron Tapley, an Earth scientist at the University of Texas at Austin's Center for Space Research. "But now static measurements are not enough."

Tapley says that the satellites will be able to pinpoint the gravitational field of areas just a few hundred kilometres across. He expects the measurements to be so accurate that even pro-



Water world: high-resolution GRACE data could allow floods in the Mississippi basin (main image) or melting ice (inset) to be monitored.

longed periods of heavy rain could turn up in the signal. "If we fly over the Mississippi, Congo or Ganges basins then we may be able to see the effects of major floods," Tapley adds.

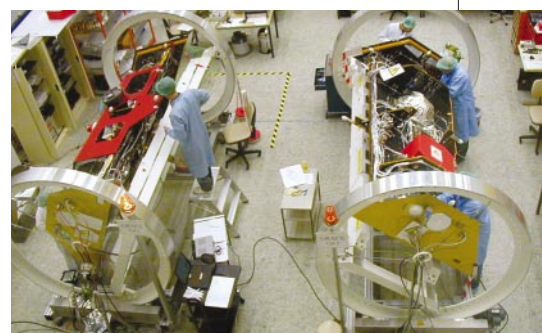
This exquisite sensitivity will allow researchers to keep track of all sorts of water movements, from changes in aquifer levels to the melting and freezing of the polar ice caps. Gravity measurements have an advantage over standard measurements of sea level or polar ice volume, in that they can distinguish expansion due to heating from an influx of water, Watkins says. "If you just measure the height of an ocean surface with an altimeter it's hard to separate a change in volume from an increase in mass."

Moisture maps

GRACE should also help scientists to track global changes in soil moisture, an important factor in the planet's hydrological cycle that influences climate change, erosion, reservoir management, water quality and ecology. Currently, researchers measure soil moisture with remote sensing in some areas, but experts agree that there is a huge need for more global coverage of the sort that GRACE could provide. Other researchers hope that GRACE will reveal density variations within the Earth's mantle, the thick layer of molten rock beneath the crust that is still shrouded in mystery.

But there are limits to GRACE's abilities. Anything that changes or cycles in less than a month, the time that it takes the satellites to complete a single scan, will be impossible to study in detail. Any change in mass smaller than about 2.8 billion tonnes will be invisible. And GRACE picks up an overall gravity signal, which is a composite of the gravitational pull from various features such as water, ice and rocks. The challenge will be to distinguish between the different materials in the signal.

To do this, mission scientists are developing algorithms that will query the data and help them to tell whether a measured shift in



State of GRACE: workers in Plesetsk prepare the twin satellites for their launch later this month.

gravity is caused by swelling water levels in an aquifer or the deeper movement of molten rocks. One factor will be the time-scale involved. Rivers would be expected to produce faster changes than ocean currents, and ocean currents should produce faster changes than deep magma flows. Other information, such as rainfall data, will help to distinguish between different events.

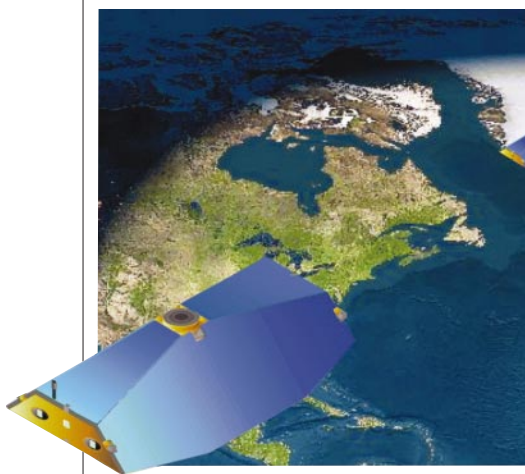
Military scientists are also likely to find applications for the data, Tapley says. For example, accurate measurements of gravity are essential for reliable guidance of long-range missiles. And the oil exploration industry is taking a keen interest, as it believes that precise gravity measurements could help to reveal the locations of untapped oil reserves.

But all good things must come to an end. The satellites will fly at a relatively low altitude of some 500 kilometres for optimal sensitivity. Here they will encounter occasional wisps of atmosphere. The satellites have accelerometers on board to sense any resulting changes in speed and so avoid distorting the data, but their orbit will gradually decay. After about five years, the very force that GRACE was designed to measure will bring the mission to a close, when the Earth's pull finally causes the satellites to burn up in the atmosphere.

David Adam is a news and features writer for *Nature*.

♦ www.csr.utexas.edu/grace

♦ op.gfz-potsdam.de/grace/index_GRACE.html



Dynamic duo: the GRACE satellites will supply regularly updated gravity maps.

L. GUMLEY/UNIV. WISCONSIN-MADISON; B. EDMISTER/SPL (INSET)

Music, maestro, please!

Neuroscientists are starting to discover how our brains process music. This involves distinct neural circuits, but a mystery remains. Why are melody, harmony and rhythm so important to us? Alison Abbott tunes in.

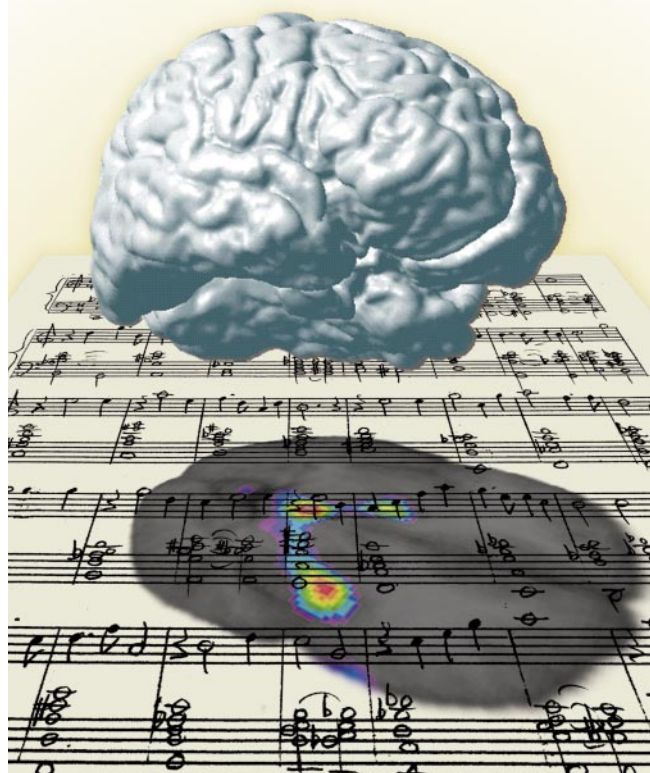
Music stirs the soul. Mozart's *Requiem* has moved listeners for centuries. And today's masters of popular culture have made a fine art of using music to manipulate our moods: when Oliver Stone picked Samuel Barber's *Adagio for Strings* as the theme for his Vietnam epic *Platoon*, he knew that its plaintive tones would induce a surge of emotion among movie-goers.

Now, research is revealing the neural interactions that provoke these reactions, showing that our brains have distinct circuits for perceiving, processing and playing music. The neuroscientists behind these studies say that music provides an ideal model for studying how the brain integrates complex perceptual and behavioural tasks. But these researchers are also interested in answering a more fundamental question: what is music for?

After all, an appreciation of music confers no glaringly obvious advantage in the darwinian struggle for survival. Various theories have been put forward — that music promotes social cohesion, for example — but so far, none represents more than a plausible 'just so' story. Our love of music might merely be a pleasurable side-effect of the evolution of other perceptual abilities — representing, as cognitive neuroscientist Steven Pinker of the Massachusetts Institute of Technology has put it, "auditory cheesecake"¹.

Each of our sensory organs delivers its information directly to its own specialized area within the temporal lobes of the cerebral cortex. These areas are divided into three hierarchical levels — primary, secondary and tertiary — according to the complexity of the data-processing task. Together, they build up our mental representations of the world.

But responding to a piece of music requires far more than the perceptual skills of the auditory cortex, a structure sitting in each of the brain's temporal lobes within a



Thrills and chills: music can provoke an intense emotional response, but how — and why — do our brains react so dramatically to musical stimuli?



fold of tissue known as the lateral sulcus (see figure, opposite). For one thing, it requires a working memory, where snatches of music can be stored while the brain makes sense of the whole tune. This memory is likely to be located elsewhere in the temporal lobes and towards the rear of the frontal lobes², but scientists have not yet pinned this down. Where long-term memories for tunes are stored is even less clear, but they are remarkably resilient: once we have learnt a melody we rarely forget it. And then there is the emotional response to music, which is processed in a set of structures distributed more widely through the brain.

Melody makers

Making music, as opposed to just listening to it, requires well-honed motor skills, and relies on a high level of integration between auditory inputs and motor control. The fine hand, finger or lip control required to play an instrument are, like all movements, directed by specialized motor areas of the cerebral cortex. Although detailed studies of the brain areas involved in musical dexterity have yet to be conducted, the secondary motor cortex directs more complex motor patterns, such as coordinated finger move-

ments, whereas the primary motor complex generally handles simpler tasks.

Playing an instrument also depends on the brain interpreting somatosensory — touch — information from the fingers and lips in contact with the instrument. "The level of motor and sensory processing that goes on in playing and responding to music, in addition to auditory processing, is extraordinary," says Eckart Altenmüller, director of the Institute of Music Physiology and Musician's Medicine in Hanover, Germany. "It is the perfect paradigm of complex human behaviour." In other words, neurological studies of music processing can serve as a model for studying how the brain works more generally.

But such studies have lagged behind those of visual perception. In the 1960s, neuroscientists began to expose animals to features such as patterns of lines, colours and edges, and recorded the subsequent electrical activity in the subjects' brains. From this they could map the precise areas within the visual cortex in which such features are perceived. But these studies are difficult for music, which can only be examined in human volunteers, where it is not possible to place electrodes deep in the brain. What is more,

notes played disappear the moment they are perceived.

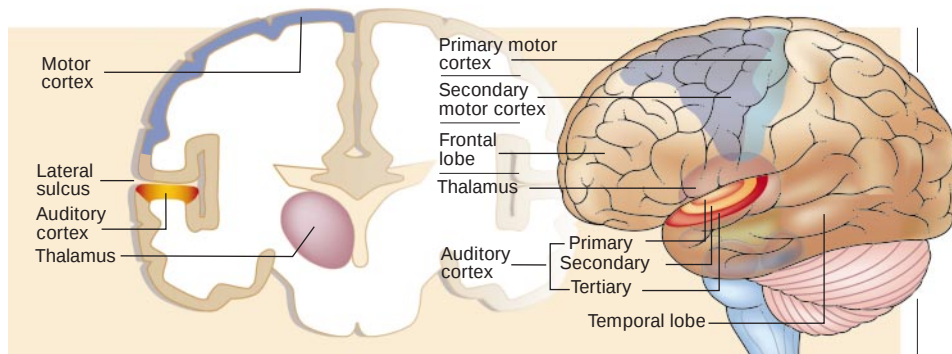
"This is why it has been hard to get a handle on music," says Robert Zatorre, director of the Montreal Neurological Institute and Hospital at McGill University in Canada, who began investigating the neurobiology of music two decades ago. He found himself more or less alone in the field until the mid-1990s, when others entered the fray. Like Zatorre, who is an organist, most of these researchers are themselves musicians.

What we now know about the neural circuits that process music has mostly been determined by using the brain imaging techniques of positron emission tomography (PET) and functional magnetic resonance imaging, which can determine the anatomical location of activity within the brain. Other methods include electroencephalography and magnetoencephalography (MEG), which respectively measure electrical or electromagnetic activity of neurons within the cerebral cortex that lie close to the skull. Musical deficits in brain-damaged patients — those who have suffered a stroke or have had all or part of their temporal lobes removed to treat epilepsy — have also yielded insights.

Three-part harmony

Using these techniques, neuroscientists have studied activity in the auditory cortex in response to music³. Although the distinctions between the anatomical location of different levels of processing have not been as well defined as for vision, the primary cortex, which receives input from the ear from a structure called the thalamus, is mainly thought to identify the fundamental elements of music, such as pitch and loudness. The secondary cortex is believed mainly to focus on harmonic, melodic and rhythmic patterns, and the tertiary auditory cortex is thought to integrate these patterns into an overall perception of the music.

Studies have also revealed that music is processed by specific neural circuits that might overlap with language and other higher functions, but are certainly not subservient to them. For example, the higher levels of music perception are, like many



Musical information is relayed from the thalamus to the auditory cortex, which is split either side of the brain. To play, musicians must integrate this information with signals from the motor cortex.

sophisticated cognitive tasks, to some extent 'lateralized' — that is, they take place predominantly in one side of the brain. Language is predominantly processed in the left hemisphere, along with a variety of reasoning tasks, but in most people the higher levels of musical perception take place predominantly in the right hemisphere⁴, where the processing of emotional and spatial information takes place. There is also evidence of some lateralization at the level of the auditory cortex³.



Isabelle Peretz sees tone deafness as a highly specific neural disorder.

with congenital amusia — commonly known as 'tone deafness', whose sufferers have reportedly included Theodore Roosevelt and Che Guevara.

None of Peretz's subjects noticed when a note in a tune was changed by a semitone. And they were unmoved by dissonance, combinations of notes that sound unpleasant to most people. But the volunteers could all recognize the pitch changes that convert a statement, such as 'He speaks French', into a question: 'He speaks French?'

In addition, most of Peretz's amusical subjects could neither be taught nor remember a melody, and most could not tap in time to a tune. "It is an extraordinarily specific disorder, confined to the musical domain," says Peretz. "We think that congenital amusia results from some disruption of the wiring in the auditory cortex, but we'll have to do more detailed functional neuroimaging studies to understand exactly what has gone wrong."

The neural circuits associated with music can run independently of input. Everyone sings to themselves — and who hasn't experienced the irritating phenomenon of being unable to get a tune out of their

head? Andrea Halpern, an experienced singer and a cognitive psychologist at Bucknell University in Lewisburg, Pennsylvania, has used PET to observe the brains of subjects as they imagined a specified melody. She found that the areas activated were similar to those that responded when music was played — except for the primary auditory cortex, which remained quiet⁷. "The signals are weaker, but the music is played in real time, and subjects react in much the same way as they do when listening to music," says Halpern.

With this basic knowledge to hand, neuroscientists are starting to use studies of music processing to address some of the big questions in brain research. One is the extraordinary plasticity of the cerebral cortex, which allows us to learn new skills. Using MEG, for instance, Christo Pantev of the University of Münster in Germany and his colleagues showed in 1998 that a much larger area of the auditory cortex was activated in musicians on hearing a piano tone compared with non-musicians. The earlier the age at which the musicians had begun to practise, the greater the enlargement⁸.

Now at the Rotman Research Institute in Toronto, Canada, Pantev has shown that the training effect depends on the instrument involved. Last year, in a study of trained violinists and trumpeters, he saw that the enlarged area of activity in the auditory cortex was smaller when the violinists listened to trumpet tones, and vice versa, than when each listened to the sound of their own instrument⁹.

Senses working overtime

In other experiments, conducted with Thomas Elbert of the University of Constance in Germany, Pantev has shown that skilled violinists give over a larger area of their somatosensory cortex to represent each of the four fingers used to hold down their instruments' strings than do non-musicians. The precise location of this region was also shifted compared with non-musicians, and again, the magnitude of these changes correlated with the age at which the violinists had started playing¹⁰. In trained musicians, the area of the motor



All in the mind: Robert Zatorre has pioneered research into the neurobiology of music.

cortex governing hand movements also seems to be enlarged¹¹.

Trained musicians also tend to use more of the left half of their brains for music processing, probably because they have learnt to process musical information more analytically⁴. But it may be possible to over-train. Altenmüller, who is a flautist, believes that excessive practice is the cause of the focal dystonias — which can cause two fingers to move as if they are joined together — that affect around one in a hundred musicians. In a study of eight sufferers, Altenmüller and his colleagues found that the representations of the two fingers in the somatosensory cortex had enlarged so much that they had merged¹². “This may be reflected in merged movement, and may indicate a biological limit to how far we can develop our musical skills,” says Altenmüller.

Although training undoubtedly rewires the cerebral cortex, it is still unclear whether training alone can create talented musicians, or whether it just makes inherently talented people better.

Hothouse flowers

To investigate this further, Pantev is now embarking on a study of children at the Suzuki School of Music in Toronto and Hamilton. This is a hothouse conservatory, which takes children from the age of four. Pantev’s team will study the anatomical location of the children’s brain activity in response to music, as well as their perceptual and cognitive skills, before they start training, and follow their development over the years. This will be a decisive study, says Zatorre, who suspects that it will reveal musical ability to be a “bit of nature, a bit of nurture, like everything else”.

Evidence that musical ability has some heritable component came last year, from a study of 136 pairs of identical twins and 148 pairs of non-identical twins. Within pairs, the performance of the identical twins in a task of identifying melodies with an incorrect note was more closely matched than for the non-identical twins¹³.

But the deepest mystery is why music is so emotive. Most of us have experienced the



Christo Pantev is using magnetoencephalography (inset) to track neural development in young musicians taught by the Suzuki method.

‘chills’, when a piece of music sends shivers down your spine and raises the hair on your skin. “When you ask people why they listen to music, they invariably refer to some emotion,” says Zatorre. “But emotion has been ignored in music research until recently, because it is so very hard to study.”

Zatorre has led the way. Initially, he focused on negative emotion, working with Anne Blood, then also at the Montreal Neurological Institute and Hospital. “Musical tastes are so personal that it was easier to design an experiment involving the sort of musical feature everyone dislikes — dissonance — rather than finding what everyone will respond to positively,” says Zatorre. His and Blood’s PET study on 10 volunteers showed different patterns of activation in specific brain areas known to be involved in emotional processing, in response to consonant and dissonant melodies¹⁴. These areas are separate from the auditory cortex, indicating that the recognition of music and its emotional appreciation use different neural pathways.

Hit parade

But Zatorre and Blood, who is now at the Massachusetts General Hospital in Charlestown, are most interested in understanding the positive side of musical emotion. In a more recent PET study, 10 music students listened to pieces of music they had individually chosen for their powerful, but very personal, emotional impact. The pieces activated neural pathways similar to those associated with euphoria and reward, which are also activated in response to other pleasurable activities — eating and sex, for example¹⁵.

Although eating and sex are directly related to survival and reproduction, it remains unclear why evolution should have shaped our brains to derive pleasure from music. Most explanations argue that music plays a key role in social cohesion, or that it promotes the development of motor and perceptual skills that are crucial to

survival. More fancifully, some scientists have even suggested that a mother’s singing can cause her baby’s hair to stand on end and so keep it warm.

Whatever the function, many of the neuroscientists studying music are convinced that it is the product of evolutionary design. “The emotional processing of music probably does have an evolutionary basis,” says Altenmüller. “The chills can be so profound, so biological, it seems very likely that the evolutionary point is about social cohesion.”

Pinker remains unconvinced, arguing that our appreciation of music and our musicianship are by-products of the way our auditory system analyses sounds, combined with our tool-making skills. “The problem with most of the theories put forward is that they are circular,” he says.

The true answer may never be known. “We can’t do the experiment to re-evolve ourselves,” observes Halpern. Thankfully, we don’t need to understand the evolutionary significance of music, nor the details of how it is processed in the brain, for it to continue to work its magic. As Elvis Presley once said: “I don’t know anything about music. In my line you don’t have to.”

Alison Abbott is Nature’s senior European correspondent.

1. Pinker, S. *How the Mind Works* (Norton, New York, 1997).
2. Zatorre, R. J., Evans, A. C. & Meyer, E. J. *J. Neurosci.* **14**, 1908–1919 (1994).
3. Zatorre, R. J., Belin, P. & Penhune, V. B. *Trends Cogn. Sci.* **6**, 37–46 (2002).
4. Peretz, I. in *Music and Emotion: Theory and Research* (eds Juslin, P. & Sloboda, J.) 105–134 (Oxford Univ. Press, 2001).
5. Ayotte, J., Peretz, I. & Hyde, K. *Brain* **125**, 238–251 (2002).
6. Peretz, I. et al. *Neuron* **33**, 185–191 (2002).
7. Zatorre, R. J., Halpern, A. R., Perry, D. W., Meyer, E. & Evans, A. C. *J. Cogn. Neurosci.* **8**, 29–46 (1996).
8. Pantev, C. et al. *Nature* **392**, 811–814 (1998).
9. Pantev, C., Roberts, L. E., Schulz, M., Engelien, A. & Ross, B. *NeuroReport* **12**, 169–174 (2001).
10. Elbert, T., Pantev, C., Wienbruch, C., Rockstroh, B. & Taub, E. *Science* **270**, 305–307 (1995).
11. Amunts, K. et al. *Hum. Brain Mapping* **5**, 206–215 (1997).
12. Elbert, T. et al. *NeuroReport* **16**, 3571–3575 (1998).
13. Drayna, D., Manichaikul, A., de Lange, M., Snieder, H. & Spector, T. *Science* **291**, 1969–1972 (2001).
14. Blood, A. J., Zatorre, R. J., Bermudez, P. & Evans, A. C. *Nature Neurosci.* **2**, 382–387 (1999).
15. Blood, A. J. & Zatorre, R. J. *Proc. Natl Acad. Sci. USA* **98**, 11818–11823 (2001).



Same old song: music has played an important role for millennia. Instruments similar to this bone flute have been found at Neanderthal sites.

Curtain has fallen on hopes of legal bioprospecting

Local communities, too, could have benefited from better health care and conservation.

Sir— Your News Feature “The curtain falls” (*Nature* **414**, 685; 2001) about the termination of the International Cooperative Biodiversity Group (ICBG) in Chiapas, Mexico, was an excellent analysis of the complex issues that brought this innovative research and development project to an unfortunate end. However, there are three misinterpretations in the article which we, as members of the US Interagency Technical Advisory Group to the ICBGs, would like to correct.

First, you state that the aim of an ICBG is to find natural products for treating important diseases in the United States. This multifaceted programme, supported by several US government agencies, actually has the goal of identifying potential drugs to enhance the public health of both developing and developed countries, while promoting economic development and conservation of local biodiversity. Each ICBG is required to focus on disease areas of local importance as well as those of importance elsewhere.

The Chiapas ICBG was no exception, and planned to research potential treatments for diarrhoea, respiratory conditions, infectious diseases, contraception and other locally important health needs. Substantial effort and investment is made by all ICBGs to help local organizations develop sustainable economic uses of their natural resources, as well as models for sharing profits and other benefits that emerge from collaborative research efforts.

Second, referring to the innovative theatre used by the Chiapas ICBG for developing informed consent in Maya communities, your article’s standfirst asked “But did the plays distract attention from the involvement of commercial interests?”. This could give the mistaken impression that distracting the audience from the possibility of commercial development was intentional or desirable. The participation of private companies to develop and market therapeutic agents is detailed in every publication or presentation about the ICBG programme, and was an explicit component of this community theatre.

Finally, you state that the National Institutes of Health asked the Chiapas principal investigator, Brent Berlin, to suspend collecting plant material late last year. In fact, the ICBG never initiated collections for drug discovery. Pressed, fumigated botanical specimens for taxonomic museum reference were collected in the first year of the project under a scientific collection permit issued

by the Mexican government. When that permit expired, the group suspended these collections. The investigators subsequently decided that it was unproductive to pursue federally permitted taxonomic collections because these had provoked concern among people who were confusing basic research for taxonomy with potentially commercial drug discovery.

The controversy that has led to the termination of the Chiapas ICBG may have a chilling effect on the ability of scientists to develop transparent and ethical collaborations in natural-products drug discovery, biotechnology and other sustainable uses of biodiversity for local and global benefit. In our opinion, all parties have lost, not least local

communities in developing countries. These stand to benefit from improvements in health care and from enhanced capability to use and conserve their disappearing biological resources and associated traditional knowledge.

Joshua Rosenthal

Division of International Training and Research, Fogarty International Center, National Institutes of Health, 31 Center Drive, Bethesda, Maryland 20892, USA

Other signatories to this letter:

Flora Katz, Gordon Cragg, Yali Hallock, George Johnson, Linda Brady, Michael Gottlieb, Chris Tseng, Richard Hawks, Jamie Biswas National Institutes of Health, Bethesda, Maryland, USA
Joann Roskoski, James Rodman, Elizabeth Lyons, Michael Willig National Science Foundation, Arlington, Virginia, USA
Carol Kramer-LeBlanc, Anne Bertinussen US Department of Agriculture, Washington DC, USA

Don't fight fire with fire

Sir— The fires that ringed Sydney in December and January were as intense as any in the past 70 years (see *Nature* **415**, 105; 2002). Over half a million hectares of bush land burnt and more than 10 times this area threatened but saved; about 100 houses were destroyed. The fires have stimulated strident demands for more frequent and extensive burning off (‘hazard reduction burning’), especially in national parks, to protect property.

Most authorities agree, though, that protection is not guaranteed by this approach. The way forward is not simple. It must include strategic (not broad-scale) hazard reduction, continued emphasis on fire-fighting, education and better town planning in areas of high fire risk.

The urban areas of Sydney and Wollongong are surrounded by large conservation areas: a World Heritage Area, several national parks and water catchments. The principal objective of the New South Wales National Parks & Wildlife Service, the agency responsible for most of this land, is “to protect and conserve natural and cultural heritage”, including natural ecosystems, biodiversity in general, and species and communities listed as vulnerable and endangered. Although the proximity of the threatened conservation areas to Sydney makes the city stunningly beautiful, the variegated boundaries create an unenviable challenge for land managers, who have to protect the native biota inside the boundaries from big fires while protecting people outside.

Many plant and animal species are threatened by too-frequent fires, even though they evolved with fire as a natural

disturbance. Paradoxically, some plants most favoured by fire are actually killed by it. They re-emerge, phoenix-like, from seeds that were protected from heat in woody fruits or the soil. The juvenile period — the time needed for the new recruits to develop a seed bank of their own — can exceed 10 years for some species. A second fire during this time could spell local extinction. Also, animals requiring dense habitat are confined to areas that do not burn frequently. These species need long-unburnt refuges from which to re-invade, once the burnt vegetation recovers to an appropriate stage. Frequent burning sustains unsuitable habitat. The current state of knowledge is summarized in *Flammable Australia: The Fire Regimes and Biodiversity of a Continent*, edited by R. A. Bradstock, J. E. Williams and A. M. Gill (Cambridge University Press, 2001).

The primary conservation objective of the national parks is that of managing too-frequent fires, so managers and politicians must resist the demand for broad-scale hazard reduction. Perhaps neighbouring property and lives can be protected by attention to the boundaries of conservation areas, without burning frequently throughout the parks. This strategy has been pursued in New South Wales national parks, to the extent that biodiversity conservation in these boundary zones may be sacrificed to protect property. Backed by a well trained, resourced and coordinated fire-fighting effort, this strategy meant that remarkably few houses were lost in the recent fires. Now it is time to find and defend solutions that also protect the parks.

Rob Whelan

Institute for Conservation Biology, University of Wollongong, Wollongong, NSW 2522, Australia

The book of revelation

Discomfiting last words from an icon of evolutionary biology.

The Collected Papers of W. D. Hamilton: Narrow Roads of Gene Land. Volume 2: Evolution of Sex by W. D. Hamilton
Oxford University Press: 2001. 928 pp.
£50, \$85

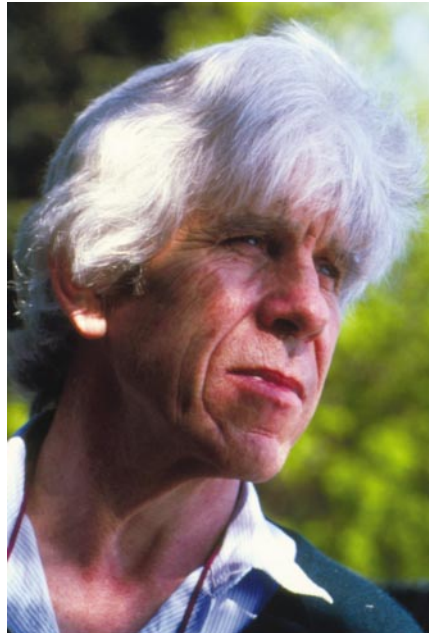
Olivia P. Judson

This book is the evolutionary biologist's *Harry Potter*: the long-awaited sequel to a captivating story of a young man with extraordinary powers. For W. D. "Bill" Hamilton was one of the twentieth century's greatest evolutionary biologists. He saw the world in a profound new way and, in elucidating his vision, changed for ever the way the rest of us see the world too. His discovery of inclusive fitness — namely, that an organism can spread its genes by helping its relations to spread theirs — blasted a path for understanding the evolution of altruism, and incidentally founded the field of sociobiology. His work on the evolution of sex ratios, which predicted when sex ratios might deviate dramatically from one-to-one, pioneered the use of game theory in evolutionary biology.

With the first volume of his collected papers (*Narrow Roads of Gene Land. Volume 1: Evolution of Social Behaviour*, W. H. Freeman, 1996), Hamilton created a dazzling new form of scientific autobiography. Each of the 15 papers he published between 1963 and 1980 is preceded by a brief sketch of his life and thoughts at the time he was working on a given idea. Funny, erudite, compelling and often poignant, these sketches show a lonely outsider struggling to make his radical ideas heard.

By itself, the brilliance of Volume 1 would have guaranteed interest in the sequel. But there is another, sadder, factor. In March 2000, Bill Hamilton died, the autobiographical sketches for Volume 2 just completed. In a real sense, then, these sketches are his last words. Mighty strange last words they are.

No longer straight-up autobiography, the sketches in this volume are a thorny and unweeded thicket of scientific advocacy, political manifesto, broken intellectual taboos, self-revelation and apocalyptic vision. They swirl from one topic to the next so fast that sometimes the reader has the impression of being party to a phantasmagoria. Hamilton does recount plenty of entertaining stories: the "Greasy Sheet Award" for the worst hotel he ever stayed in goes to a bed and breakfast in Ireland, where not only were the sheets dirty, but the landlady's son crept into the room in the middle of the night and slept on the floor —



Bill Hamilton's view of why animals help their relatives formed the basis of sociobiology.

and there wasn't any breakfast. Hamilton's erudition and humour are still there, but they are mingled with anger, frustration and bitterness.

Volume 2 opens with him working in America, at the University of Michigan. By this point, his early ideas — the papers of Volume 1 — had been accepted, and he had come in from the cold. This surely made for a more conventional, even mundane, academic life. Moreover, Hamilton had, by and large, shifted his attention to the evolution of sex, a problem he grappled with until he died. The 18 scientific papers of this volume — published between 1980 and 1991 — reflect the change in both focus and lifestyle. In Volume 1, only three papers have co-authors, and virtually every paper tackles a different subject. Here, almost half have co-authors, and more than half are devoted to exploring aspects of Hamilton's favoured explanation for the evolution of sex — this invokes antagonistic coevolutionary interactions between hosts and their harmful parasites, a theory known as the Red Queen.

It is here that the scientific advocacy comes in: as he admits in the preface, Hamilton felt (correctly) that many of the scientific ideas represented in this volume are not generally accepted, and he thus devotes considerable chunks of the sketches (as well as two appendices) to additional argument and persuasion. The anger and bitterness are also explained. Hamilton includes a citation analysis showing that one of his papers,

which was twice rejected by *Nature*, has since been more influential than two papers on the same subject (by other authors) that *Nature* accepted at roughly the same time. But although the emotions he describes will be familiar to anyone who has had a cherished paper rejected, this passage amounts to an object lesson on why it is ridiculous to become exercised over losing *Nature's* lottery. What is perhaps most startling about all this, however, is Hamilton's openness about his disappointments and set-backs. In writing autobiography, most authors polish the image they present to the world and hide moments of anguish and defeat.

But the self-revelation is nowhere near as startling — or disturbing — as the mix of politics and apocalypse that suffuses the sketches. Take the politics first. Hamilton is an unabashed, no-fig-leaf naturalist. He believes that genetics, not nurture, accounts for a large and important range of human behaviour — from racism and xenophobia to differences in intellectual abilities between men and women — and that only by admitting and understanding this, only by casting aside hypocrisy on the matter, can fundamental human problems be tackled. As an example, he argues that a basic cause (emphatically not a justification) of racism — and, particularly, of ethnically motivated genocide — is a differential birth rate between groups. And, yes, he does extend this to the Nazi extermination of Jews.

Much of this makes uncomfortable, even offensive, reading. But to be fair, much of the discomfort has less to do with the truth (or not) of Hamilton's statements than with the fact that he raises possibilities that most of us shrink from contemplating. After all, it is self-evident that the human psyche has been shaped by natural selection; and it is certainly possible that the shape is an ugly one. But I feel — and this is where I completely disagree with Hamilton — that, although understanding how natural selection has acted on humans will help us to understand why we are the way we are, it tells us nothing about what we would like to become.

And the apocalypse? For the human species, Hamilton foresees an imminent mutational meltdown — the crippling accumulation of deleterious mutations — as a consequence of modern medicine. He evaluates three courses of action: do nothing (which will lead to the coming of the "planetary hospital" and certain doom), germline intervention (desirable but impractical, perhaps impossible), and a return to natural selection (salvation). Whew! Of course, this question is worth investigating. If humans

have done away with natural selection — which is by no means obvious — then deleterious mutations will start to accumulate. Even so, Hamilton's view is inconsistent in several respects. He despises efforts to keep children with genetic defects alive, and abhors the fact that caesarian sections allow women with narrow hips to give birth. However, neither of these supposed evils seems likely to contribute significantly to mutational meltdown. Moreover, he ignores other, more important sources of relaxed natural selection — such as vaccination programmes, and an adequate food supply. And as for returning to natural selection — well, more natural selection equals more death, which is hardly something to agitate for.

The full strangeness of this book cannot easily be explained away. None of the themes is new. A close reading of Volume 1 shows that all the currents — the naturism, the concern about deleterious mutations, even some of the personal bitterness — are lurking in the background. For those who knew Hamilton well, this book evokes his voice, his manner, his laughter. I cannot agree with much of the content. But perhaps it is not surprising that someone with such an original view of the natural world should also have disconcerting and discomfiting views of *Homo sapiens*. ■

Olivia P. Judson is in the Department of Biological Sciences, Imperial College, London SW7 2AZ, UK. She is presently visiting the Institut des Sciences de l'Evolution, Université de Montpellier II, Place Eugène Bataillon, 34905 Montpellier Cedex 05, France. She was one of Hamilton's graduate students.

A revolutionary way with weirdness

Quantum Computation and Quantum Information

by Michael A. Nielsen & Isaac L. Chuang
Cambridge University Press: 2000. 700 pp.
£80, \$130 (hbk), £29.95, \$48 (pbk)

Seth Lloyd

Quantum computers are devices that store and process information at the level of individual atoms, nuclei and photons. The past decade has seen an explosion of work on quantum computation and quantum information. For theorists, quantum computation provides a new, physically motivated paradigm for information processing. For experimentalists, quantum information and computation provides a set of difficult technical challenges which, if they can be met, will result in the development of valuable quantum technologies. The field is undergoing a revolution.

But this is the second revolution in quantum mechanics and information: the first

Beautiful beetles

Beetles come in a dazzling range of colours, but few of us really appreciate their looks.

Poul Beckman hopes to change that with *Living Jewels: The Natural Design of Beetles* (Prestel, £24.95, \$39.95), a collection of full-page photographs of beauties such as *Sternotomis bohemani*, shown here.

began with the discovery of quantum mechanics itself. That the origins of quantum mechanics are tied to information might at first seem strange. After all, quantum mechanics is a century old and information theory is only half that age. But the basic formulae of information theory were developed in the latter part of the nineteenth century, in the field of statistical mechanics, and were incorporated in the quantum version of statistical mechanics early in the twentieth century. To phrase this in contemporary terms, Maxwell, Boltzmann and Gibbs realized that the statistical mechanical quantity called entropy was in fact a form of information — information about the microscopic state of a fluid, gas or solid that we do not possess. So even though the mathematical theory of information did not at that time exist, the formulae describing information were incorporated in quantum mechanics from the start.

This first revolution arose because quantum mechanics puts a limit on the amount of information — measured in bits — that can be measured in physical systems, and allows it to be quantified precisely. It resulted in the contemporary theory of statistical mechanics and solid-state physics, as well as in quantum limits to the capacity of communication channels. The second revolution arose because quantum mechanics also possesses a variety of strange and counter-intuitive features that go under the collective title of 'quantum weirdness'.

In the early 1980s, Paul Benioff showed that computers could operate in a way that preserved the coherent, or wave-like, aspects of quantum mechanics that are responsible for quantum weirdness. Then Richard Feynman and David Deutsch suggested possible ways in which this weirdness might be exploited to outdo classical computers. But the applications suggested by Feynman and Deutsch, though intriguing, were not sufficient to spur a large-scale push

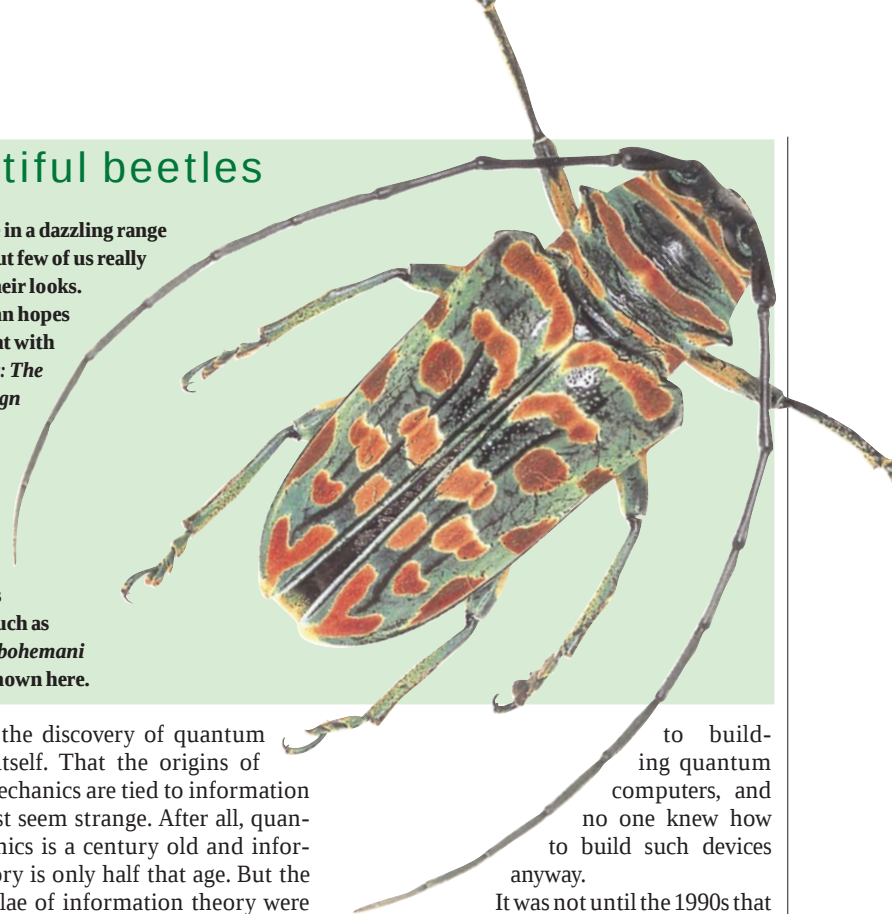
to building quantum computers, and no one knew how to build such devices anyway.

It was not until the 1990s that Peter Shor showed that quantum computers, if they could be built, could be used to factor large numbers. This is a problem with significant applications in communication and cryptography. (Essentially, the security of so-called 'public key' cryptosystems, which are used to conduct secure transactions over the Internet, relies on large numbers being hard to factor. So factoring constitutes a 'killer application' for quantum computers.)

In addition, a number of researchers showed that quantum computers could be built using existing quantum technologies, such as quantum optics and nuclear magnetic resonance. Designs were created and prototype quantum information-processing devices were constructed and operated. Quantum communication protocols were discovered, and quantum error-correcting codes were developed to protect quantum information from noise and decoherence. The second revolution of information and quantum mechanics had begun.

Revolutions need texts. For a number of years, the standard text for courses on quantum information and computation was John Preskill's excellent web-based notes (see www.theory.caltech.edu/people/preskill/ph229; now available in an expanded version with Alexei Kitaev as co-author). This book by Michael Nielsen and Isaac Chuang is more extensive and deals with a broader range of topics. It also contains more introductory material, together with exercises and problems, which are useful for readers without a background in physics.

Although the field of quantum information and computation is evolving rapidly,



this book was written just after the basic algorithmic and physical tools of the field had reached a standard form. And so the text contains essentially all of the material that might be covered in a course on quantum information.

I used it as a textbook for an introductory course in quantum computation at my institute. At the end of the term, the class was polled on various aspects of the book. The response was uniformly positive: the book received ratings of 5 out of 5 for comprehensiveness, 4 out of 5 for clarity, and 4 out of 5 for problems and exercises. Within the first week, the students had bestowed on it a nickname, "Mike 'n' Ike", after the authors' first names.

A rigorous, comprehensive text on quantum information is timely. The study of quantum information and computation represents a particularly direct route to understanding quantum mechanics. Unlike the traditional route to quantum mechanics via Schrödinger's equation and the hydrogen atom, the study of quantum information requires no calculus, merely a knowledge of complex numbers and matrix multiplication. In addition, quantum information processing gives direct access to the traditionally advanced topics of measurement of quantum systems and decoherence.

Finally, understanding information in a quantum-mechanical context is crucial for the development of the next generation of quantum technologies, such as nanoscale circuits, sensors, motors and timing devices. At the nanoscale and below, information has as important a role as energy in the functioning of such devices. *Quantum Computation and Quantum Information* is a must-read for the generation of budding quantum-mechanical engineers who will build the technologies of the future. ■

Seth Lloyd is in the Department of Quantum-Mechanical Engineering, MIT 3-160, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.

Notes on a cultural theme

Beethoven's Anvil: Music in Mind and Culture

by William Benzon

Basic Books: 2001. 352 pp. \$27.50, £19.99

David Juritz

Here's an analogy that will live with me for ever — the comparison of the finale of Beethoven's ninth symphony to a troop of baboons. This striking idea is to be found in *Beethoven's Anvil*, which describes "the workshop in which human culture was first

forged and which continues to sustain us as we evolve into the future".

William Benzon, a cognitive scientist and musician, draws on research from a wide range of disciplines. Although his preface describes the book as "an exercise in speculative engineering", with the caveat that many of his speculations "must surely be wrong", he succeeds in building a convincing, though not compelling, case.

Arguing that musicality is an adaptive feature that evolved in hominids alongside speech, Benzon initially focuses on the mechanisms by which participants in musical activity coordinate their neural states. He believes that, in such states, the brain operates in a mode similar to that of dreaming, with the attendant changes in its neurochemical processes. In this state, the reduced levels of the neuromodulators serotonin and noradrenaline inhibit the laying down of long-term memories.

The sceptic in me couldn't help observing that the author nevertheless seems to have near-perfect recall of many of his own such 'ecstatic' experiences. However, he then presents compelling evidence from, among others, psychologist Jaak Panksepp (who discovered in a delightful experiment that chickens are affected by music, showing a particular affinity for Pink Floyd), that some music has a direct effect on levels of the neuropeptide oxytocin. This hormone is known to mediate bonding between individuals and may also have an enhancing role in the laying down of new memories. If these statements are contradictory, Benzon is the first to admit that further research is required.

Starting with the observation that the hypoglossal nerve controlling the tongue reached its present size 300,000 years ago, Benzon takes up the argument proposed

by zoologist Valerius Geist that the ability to mimic animal calls while hunting is an adaptive advantage that offsets the biological expense of our sophisticated vocal apparatus. (Recall President George W. Bush's recent encounter with a pretzel.) These imitative calls would then have been incorporated into ritual activity, evolving in time into a protomusical form of communication.

Hence music as a memetic phenomenon. Using Richard Dawkins' and Susan Blackmore's concept of the meme, the behavioural or cultural equivalent of a gene, Benzon shows how musical ideas and fragments that are 'successful' tend to replicate as musical language as a whole evolves.

Next, he explores music as an agent of social cohesion in hunter-gatherer and small-scale communities. He maintains that it acts to dissipate tension or to enhance the sharing of pleasure, and allows social reprogramming to occur during rites of passage. Benzon suggests that during rituals, such as a marriage between individuals from neighbouring communities, the simultaneous changes induced in the collective brain chemistry allow relatively large groups of people to adjust quickly to changes in the status of the individuals who are marrying. He argues that groups possessing the means to maintain collective harmony in this way function more smoothly, thus securing a selective advantage.

Other sections of the book address rhythm, cognitive processes, learning processes, ethnography and some music history, with a wealth of examples as illustration.

Benzon's book covers a vast area, albeit in a rather rambling fashion. But there are weak points. *Beethoven's Anvil* drops with a resounding clang at the ridiculous analogy between a troop of baboons and Beethoven's ninth symphony. Apparently, at the opening of the finale, "distinctly different ideas [for which read pseudopods formed by young males] mill about until one of them, the Ode to Joy [yes, you guessed it, a dominant male] takes charge". And they move off. A glance at the CD's sleeve notes, or the vaguest knowledge of Beethoven's compositional process, should have disabused him of such an assertion. When, in the final chapter, 300,000 years of music-making (Benzon prefers to use the painful construct "musicking") is reduced to a diatribe about racism in contemporary America and — I kid you not — the burning issue of where 'hip-hop' should go from here, my exasperation reached a level where I feared I had developed late-onset Tourette's syndrome.

This book stimulated as intensely as it irritated me. It has great scope but a lack of focus, poorly chosen examples and some weak analysis which served to alienate rather than convince. I recommend it with caution, but not as a present for Beethoven lovers. ■

David Juritz is a concert violinist based in London, UK.



A world of music: these Ladakhi musicians at Leh Gompa in Tibet emphasize music's ubiquity.

DAVID SAMUEL ROBBINS/CORBIS

Nature and function

Yvon Le Maho

One of the main scientific challenges of the new century is to assess the impact of people on the environment, including climate, and the effect that the changes we make will have on other living organisms. The key consideration is the ability of plants and animals to adapt to these man-made changes. As well as the current priority given in biology to the 'post-genomic' approach — which aims to determine the functions of the genes identified in genomic studies — it is necessary to study how whole organisms, species and communities react to environmental constraints. Ecophysiology, the science of how human, animal and plant populations behave in relation to these constraints, will have a significant part to play.

Behavioural, population and evolutionary studies, with a few notable exceptions, do not take into account the physiological boundaries within which organisms operate. It is in conjunction with these different fields of investigation that ecophysiology, which integrates behavioural and physiological responses, finds its particular appeal. Because of the innumerable interactions between living organisms and their environments, the efforts of ecophysiologicalists must be combined with those of all biologists who study animals, populations or ecosystems. This emerging integrative biology will interact with the integrative approach of those studying genes and genomes, giving rise to 'molecular ecology' and ultimately revealing the functional responses of organisms to particular environmental cues.

If, in the future, space to grow food for the world's exploding human population becomes a limiting factor, it will be important to understand how plants can capture light

most efficiently, minimize transpirational water loss, and improve nutrient acquisition. The same is also true for ecophysiological processes in microorganisms and microinvertebrates in the soil, which are among our most valuable resources.

Ecophysiology, however, faces huge technical barriers. This is particularly true when studying wild animals, because of their mobility. Sea birds roam over the oceans, fish and marine mammals dive to great depths, and migrating birds travel across deserts and over high mountains. The minute size of some animals, such as the myriad species of small insects, is also a challenge. Some of these difficulties are being surmounted by the use of ultra-miniaturized transmitters, data loggers and cameras that can be attached to free-ranging animals, and by the use of satellite-based localization and communication systems.

Many studies have shown that animals can compensate for environmental constraints by increasing their energy expenditure. However, this increase cannot exceed certain constraints, as excessive resource use would impair body condition and reduce fitness. Energetics, the study of how animals balance their energy incomes and expenditures, involves the use of tools such as stable isotopes and techniques such as heart-rate measurement. Such methods allow us to understand how wild animals' physiological performances and behavioural strategies combine to shape their life-histories and enable them to cope with environmental conditions. In turn, this will show how they optimize this synergy, and how they anticipate critical situations.

Any new investigative technique that can minimize disturbance to animals studied in the field should be tested; this will help to

Ecophysiology

To predict the future impact of mankind on living organisms, we need to know how organisms become adapted to a rapidly changing world and determine the limitations of adaptive processes.

determine whether the methods that have been used so far, and have given us our current scientific data, risk introducing flaws. Avoiding artificial results is crucial when examining the performance of free-living animals, because we need to know whether the method used to collect the data impairs the animals' function and, if so, to what extent.

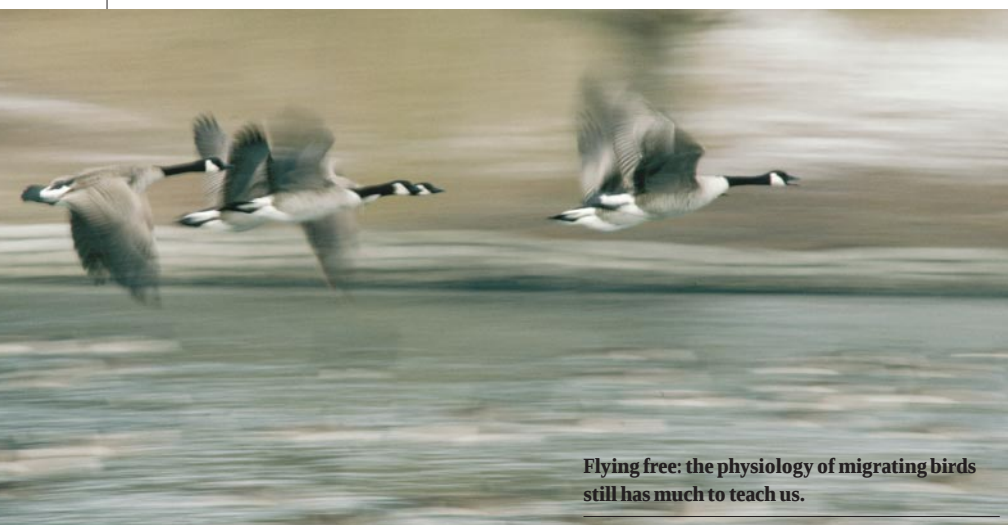
Laboratory research will remain essential to investigations of physiological processes in which conditions must be controlled — such as studies in which birds fly at steady state in a wind tunnel. However, physiological processes may be discovered in the field that could never have been found in the laboratory. For example, wild animals of many species are able to store huge amounts of fat and deplete these reserves safely, knowing when to eat before it is too late; some marine mammals dive to great depths; and incubating penguins are able to store food for weeks in their stomachs to feed the hatchling if their mate should be delayed on his return from foraging. Monarch butterflies manage their tiny reserves to travel over continents, and small passerine birds cross thousands of kilometres of desert without refuelling. Such behaviours can only be studied in the field.

Understanding processes such as these will be relevant not only to biomedicine but also to biotechnology. Indeed, considering the huge variety of conditions that animals face on our planet, their physiological adaptations represent a source of information that is still largely unexplored. The answers to the questions that their adaptations pose may well prove to be another benefit of attempts to conserve biodiversity. ■

Yvon Le Maho is at the Centre d'Ecologie et Physiologie Energétiques, Centre National de la Recherche Scientifique, 23 rue Becquerel, 67087 Strasbourg Cedex 2, France.

FURTHER READING

- Feder, M. E. *et al.* *Annu. Rev. Ecol. Syst.* **31**, 315–341 (2000).
Harvey, J. *Nature* **413**, 463 (2001).
Michener, W. K. *et al.* *Bioscience* **51**, 1018–1023 (2001).
Schmidt-Nielsen, K. *How Animals Work* (Cambridge Univ. Press, Cambridge, 1972).
Wikelski, M. & Ricklefs, R. E. *Trends Ecol. Evol.* **16**, 479–481 (2001).



Flying free: the physiology of migrating birds still has much to teach us.

LAYNE KENNEDY/CORBIS

Biodiversity equals instability?

Shahid Naeem

An understanding of how ecosystems function is vital in guiding human use of natural resources. New work will fuel further debate over the knotty problem of how an ecosystem's diversity and stability are interlinked.

The inordinate complexity of nature, whether at the level of molecular processes or of the biosphere, has puzzled biologists for some time. Does a cellular process need all those steps? Must an organism have so many genes? Does an ecosystem need all those species? Experimental biologists are currently making tremendous headway in tackling the problem of complexity, in part because of their increasing use of powerful combinatorial methods. Such approaches deconstruct biological systems into their separate parts, then systematically reconstruct arrays of replicate systems that vary in their combinations of those parts. For example, combinatorial arrays of nucleic acids and proteins can be used in genomics and proteomics studies, respectively, to better understand gene expression, metabolism and development.

Similarly, combinatorial arrays of species (Fig. 1) are becoming increasingly popular with biodiversity researchers in their quest to understand ecosystem processes such as biomass production (the amount of living matter produced), carbon cycling and nutrient retention. Combinatorial biodiversity research has been rife with debate, however¹. Progress has been made towards resolving the controversies², but the report by Pfisterer and Schmid³ on page 84 of this issue is going to upset the apple-cart.

Pfisterer and Schmid³ studied biomass production in a combinatorial plant-diversity experiment, which consisted of an array of replicate grassland plots that varied both in their number of plant species (from 1 to 32) and in their combination of species. The authors used their results to test the venerable 'insurance' hypothesis of ecosystem stability. This hypothesis is one of several that have featured in the long-standing ecological debate over the relationship between complexity (diversity) and stability⁴. Over the course of this debate, the prevailing view has see-sawed between the thesis that diversity begets stability, and the antithesis that diversity either leads to instability or is irrelevant.

Chief among the 'begets-stability' theories is the insurance hypothesis — the impeccably logical notion that having a variety of species insures an ecosystem against a range of environmental upsets. For example, suppose an ecosystem faces a drought, then a flood, which in turn is followed by a fire. According to the insurance hypothesis, if that ecosystem is

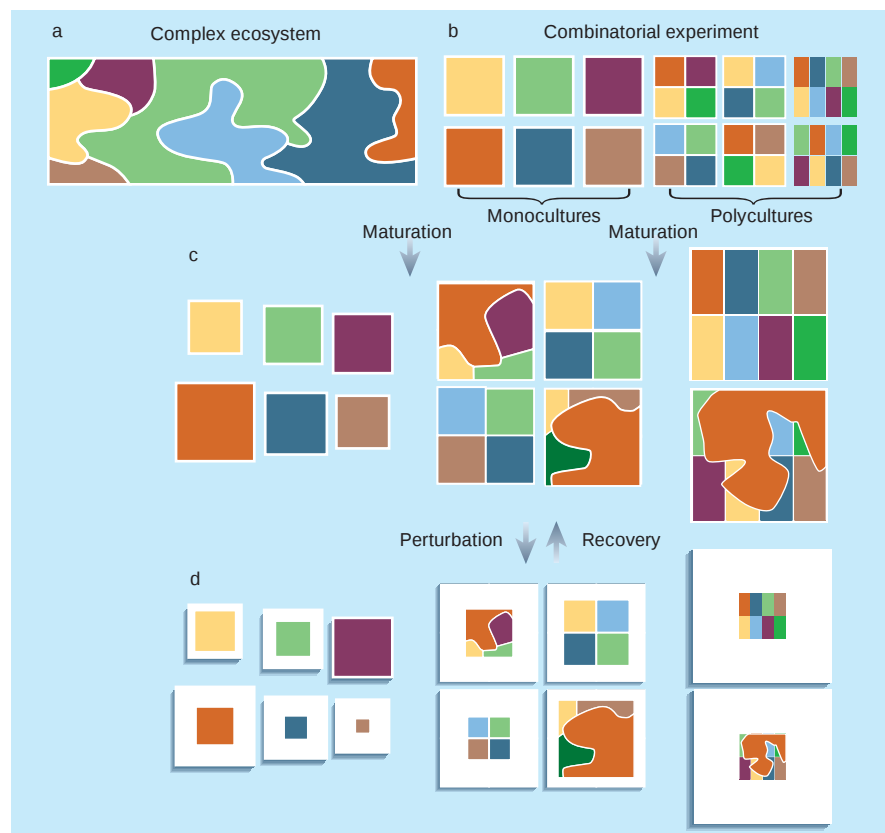


Figure 1 Combinatorial biodiversity experiments. **a**, A complex ecosystem, shown as a mixed plot of eight different plant species. **b**, Combinatorial experiments deconstruct such ecosystems into equal-sized replicates that vary in composition from monocultures to polycultures. **c**, During growth, the amount of biomass produced (represented by plot size) varies among replicates. This might be due to one of several effects, such as 'sampling' (where some species contribute more than others to the accumulating biomass; illustrated by irregularly shaped segments), or 'niche complementarity' (illustrated by most species contributing equally). **d**, Perturbations can reduce biomass (illustrated as shrinkage within frames; large reductions indicate little resistance). Recovery is driven by species growth after the perturbation; fast recovery is indicative of resilience. Combinatorial biodiversity experiments help to tease apart the relative contributions of individual species to resistance and resilience. Pfisterer and Schmid³ find that less diverse plots are more resistant and resilient than more varied communities.

diverse — if it has some species that can tolerate drought, some that are flood-resistant and some that are fire-tolerant — then two scenarios are likely. The ecosystem may show resistance, remaining broadly unchanged, because its many species buffer it against damage. Or it may show resilience: if it does get hammered, it may bounce back to its original state quickly because the tolerant species ultimately drive the recovery process and compensate for the temporary loss of their less hardy compatriots.

But Pfisterer and Schmid³ found that, when challenged with an experimentally induced drought, species-poor communi-

ties were both more resistant and more resilient (as reflected by their ability to sustain and recover pre-drought biomass production) than plots of higher diversity. The higher-diversity plots were originally more productive, but their resistance and resilience — that is, their stability — was low (Fig. 1). This is the opposite of what the insurance hypothesis predicts. It also contrasts with what combinatorial 'microcosm' experiments have found^{5,6} and what theoretical models of biodiversity have claimed⁴.

Pfisterer and Schmid's findings³ appear to support those who claim that diversity does

not lead to stability. But there's a twist, and those on each side of the debate run the risk of having their own pet theories turned against them. Pfisterer and Schmid suggest that the observed inverse association between diversity and stability is due to a theoretical mechanism known as niche complementarity. This mechanism, however, is the very same as that touted as the chief cause of the positive biodiversity–productivity relationships found in other combinatorial biodiversity experiments, such as those at Cedar Creek⁷ and those run by the BIODEPTH consortium⁸.

The central idea of niche complementarity is that a community of species whose niches complement one another is more efficient in its use of resources than an equivalent set of monocultures. For example, a uniform mixture of early- and late-season plants and shallow- and deep-rooting plants that are spread over 4 m² will yield more biomass than combined 1-m² monocultures of each species^{7,9}. So niche complementarity can explain why higher diversity tends to lead to higher productivity, and has also been adopted by those in the 'diversity leads to stability' camp because one would expect that more efficient communities would fare better in the face of stress.

Those on the other side, however, feel that existing data better support a mechanism known as sampling, where diverse communities produce more biomass simply because they are more likely to contain productive species^{10,11}. In other words, we can't read too much into experiments in which higher diversity leads to greater productivity.

What Pfisterer and Schmid suggest is that complementarity among species in a diverse plot could be its downfall when faced with perturbation. Niche complementarity is disrupted and so the whole community suffers. But this is not a problem for less diverse plots. So those in the 'diversity begets stability' camp risk being hoist on the petard of their own theory of niche complementarity. Meanwhile, although Pfisterer and Schmid's findings support the idea that diversity does not lead to stability, the authors reject a large role for sampling—the theory generally favoured by the camp that disagrees with the idea that biodiversity leads to stability.

As with other combinatorial studies, there are some limits to how widely Pfisterer and Schmid's findings can be applied. Although the authors provide evidence of niche complementarity, some contributions from sampling cannot be ruled out. Furthermore, the experiment is small-scale, short-term, does not include other contributors to the food web such as herbivores or decomposers, and looks only at drought out of a host of other possible stresses. These issues of scale and complexity, however, are faced by all combinatorial studies, including genomics and proteomics, where experi-

mental arrays are far removed from the complex systems being studied. The main point of Pfisterer and Schmid's study is that the relationship between diversity and stability may be determined by the pre-stress relationship between diversity and productivity, and the mechanisms that drive it.

Until quite recently, the contributions of living organisms to ecosystem functioning have been explored only by poking (small experiments) and peeking (observational studies). Combinatorial biodiversity research offers another approach. Some consensus will be needed before the results of combinatorial studies can be embraced with confidence. But Pfisterer and Schmid's report³ provides insights worth considering, regardless of where it stands or how it fares in the debates on how ecosystem diversity relates to ecosystem functioning and stability. The authors point out that the benefits of having productive ecosystems that provide natural resources for humans must be coun-

terbalanced by the effects that diversity has on stability. Strategies for conserving the natural world are likely to be most robust when they balance ecosystem performance with stability, along with the many other reasons for conserving biodiversity¹².

Shahid Naeem is in the Department of Zoology, University of Washington, 24 Kincaid Hall, Seattle, Washington 98195, USA.

e-mail: naeems@u.washington.edu

1. Kaiser, J. *Science* **289**, 1282–1283 (2000).
2. Loreau, M. et al. *Science* **294**, 806–808 (2001).
3. Pfisterer, A. B. & Schmid, B. *Nature* **416**, 84–86 (2002).
4. McCann, K. S. *Nature* **405**, 228–233 (2000).
5. Naeem, S. & Li, S. *Nature* **390**, 507–509 (1997).
6. McGrady-Steed, J., Harris, P. M. & Morin, P. J. *Nature* **390**, 162–165 (1997).
7. Tilman, D. et al. *Science* **294**, 843–845 (2001).
8. Hector, A. et al. *Science* **286**, 1123–1127 (1999).
9. Loreau, M. & Hector, A. *Nature* **412**, 72–76 (2001).
10. Huston, M. A. *Oecologia* **110**, 449–460 (1997).
11. Wardle, D. A. et al. *Bull. Ecol. Soc. Am.* **81**, 235–239 (2000).
12. Kunin, W. E. & Lawton, J. H. in *Biodiversity: A Biology of Numbers and Differences* (ed. Gaston, K.) 283–308 (Blackwell, Oxford, 1996).

Applied physics

Spin spotting

Hari C. Manoharan

Being able to measure electronic spin is essential for future technologies such as 'spintronics'. The combination of two imaging methods for effective spin-detection has now been demonstrated.

A visit to any hospital shows how magnetic resonance imaging (MRI) has revolutionized clinical medicine. Indeed, a manifold of resonant-imaging techniques now permeate physics, chemistry, biology, medicine, engineering and even computer science. For example, nuclear magnetic resonance, also known as MRI, can provide routine non-invasive imaging down to millimetre scales. In the laboratory, state-of-the-art MRI measurements can push this resolution down to about 10 μm. Meanwhile, the perhaps less familiar technique of electron spin resonance (ESR) has long been exploited to characterize a variety of materials.

In *Applied Physics Letters*, Durkan and Welland¹ report the successful marriage of ESR techniques to a newer technique—scanning tunnelling microscopy—to detect radio-frequency spin signals from clusters of a few organic molecules. In doing so, they combine the chemical sensitivity of spin resonance techniques with the unrivalled spatial resolution of the scanning tunnelling microscope. This result, together with the pioneering efforts of others, opens the door to a new class of studies on, or below, the nanometre scale.

At the heart of the experimental tools is the unique interaction between an applied magnetic field and the electrons and nuclei of which all atoms are comprised. Most common high-resolution imaging techniques (for

example, electron-beam microscopy) are sensitive to electronic charge, but newer, more advanced techniques rely on sensitivity to another quantum degree of freedom—spin. Progress here has been driven by both science and technology. Spin leads to magnetism, which underpins the entire magnetic-storage industry; it takes quite sophisticated techniques to image a modern-day computer hard disk in order to map out and refine the stored magnetic patterns. Looking to the future, spin could provide the basis for new computing technologies based on quantum mechanics—efforts along these lines have spawned the fields of 'spintronics' and quantum computation. But right now, the most common type of measurements that routinely access spin are spin resonance techniques.

A simple physical analogy helps to understand the basic phenomenon. Imagine a rapidly spinning top on a clean tabletop. Under the influence of gravity, the axis around which the top spins will slowly begin to rotate in small circles around the vertical. This precession, well known from playgrounds or physics demonstrations, occurs because gravity exerts a torque on the top. Because of its rotational motion, the top precesses at a known frequency determined by, among other things, the gravitational force and the spin rate of the top. Over time, because of friction, the spin axis of the top

traces out progressively larger circles as it spirals down to the table surface. If you were to measure the precession rate, you would learn something about the structure of the top — for example, its mass and geometry — and also about its local environment.

Remarkably, this analogy can be translated to the quantum world to gain insight into the behaviour of particle spins. An applied magnetic field takes on the role of gravity in the spinning-top analogy. Both electrons and nuclei possess spin, and these begin to precess around the direction defined by the field. The frequency of precession scales with the applied magnetic field, and is about 1,000 times faster for electrons than for nuclei.

MRI techniques work by applying and measuring time-varying signals that are in resonance with the spin-precession frequencies of atomic nuclei and hence sensitive to the local environment and chemical species present. But the spatial resolution of MRI is limited in part because it takes around 10^{15} nuclei to generate a detectable signal. Even though the precession frequencies of spins in conventional ESR techniques are higher and so are easier to detect, about 10^{10} electrons are still required for a measurable signal.

If the number of spins needed for detection can be scaled down, the spatial resolution automatically goes up. How far can this scaling go? Surprisingly, there is no known physical principle that precludes the logical end point of this downscaling — measuring a single spin. Although other techniques already provide access to single atoms and molecules, this experimental feat would add the benefit of 'seeing' an important attribute — the elusive spin degree of freedom to which other imaging methods are blind.

New techniques are making fantastic strides along these lines. After considerable effort, magnetic resonance force microscopy² (which exploits microfabricated cantilevers to detect the tiny forces emanating from precessing spins in a magnetic field) has recently achieved sensitivity to about 100 electron spins^{3,4}. Several groups are hard at work in pushing detectability down to a single electron spin; then the target would be the even more elusive nuclear-spin signal. These groups have their work cut out for them.

In contrast, the scanning tunnelling microscope (STM) is a natural platform on which to base spin-sensitive imaging, as it already provides routine access to single atoms and molecules. Exploiting quantum mechanics, this microscope works through the action of electrons 'tunnelling' across a normally impenetrable barrier — here, the gap between the sharp metallic tip of the microscope and a flat sample. When the gap size is just a few atomic diameters, electrons can magically transport themselves across the gap because the electron clouds emanat-

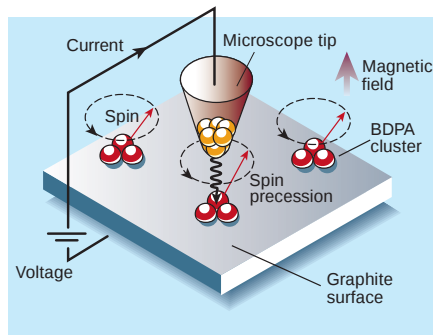


Figure 1 Spin detection through the union of high-resolution microscopy and resonance techniques. Durkan and Welland¹ coated a graphite surface with clusters of organic BDPA molecules. In the applied magnetic field, the electron-spin vectors (red arrows) associated with free radicals in the molecules precess at a certain frequency — just as a child's spinning-top precesses in a gravitational field. When the tip of a scanning tunnelling microscope is brought close to (but not touching) a cluster, a current flows between the tip and the sample (wavy arrow), due to the quantum-mechanical 'tunnelling' effect. This current is modulated at the precession frequency; detecting the modulation effectively measures electronic spin in the molecule.

ing from the tip and the sample begin to overlap. This 'tunnel current' is an extremely sensitive function of the gap size. By recording the current variations as the tip scans the sample, the surface can be mapped out down to the level of single atoms.

Only recently has the function of the STM been successfully extended to spin-sensitive imaging. The STM has been used to image local structure in magnetic materials^{5,6}, and a few experiments have provided access to single spins or magnetic moments by detecting the tiny perturbation these spins cause in a metallic^{7–9} or superconducting¹⁰ substrate. But these studies were all static or quasi-static in nature. In principle, spin precession, leading to a time variation in the microscope tunnel current, could be harnessed for even more sophisticated measurements.

The first pioneering attempt to detect this time-varying ESR component with STM was performed over a decade ago by Manassen and colleagues¹¹. Their results were controversial for several reasons: the spin resonance signals were quite small and buried in noise; furthermore, they emanated from a then unknown defect in silicon; and although it was easy to speculate, there was not an adequate theory to explain why a precessing spin should modulate the tunnel current. Manassen has subsequently improved this original work¹², and new theory has since emerged to explain the effect¹³. Putting the final nail in the coffin of controversy, Durkan and Welland¹ have now performed a convincing ESR–STM experiment on the organic molecule known as BDPA. The free

radicals in this compound generate a large spin signal that has been well characterized in conventional ESR studies.

To perform the experiment, Durkan and Welland coated a graphite substrate with a solution of BDPA and then dried the solvent, leaving a distribution of BDPA molecules in small clusters across the surface (Fig. 1). Applying a magnetic field caused the free-electron spins in the molecules to precess at a known frequency, called the Larmor frequency. With the STM tip above the coated surface, the authors measured a radio-frequency modulation (at the Larmor frequency) of the tunnelling current. They found that the proportionality constant (the 'g-factor') between the applied magnetic field and the Larmor frequency was very nearly equal to 2 — conventional ESR experiments confirm this g-factor for BDPA molecules. The authors also demonstrate that the observed precession frequency scales linearly with applied magnetic field, and disappears over the bare graphite surface.

The time-dependent radio-frequency signals appeared and disappeared on millisecond timescales. Presumably, although the individual molecules could not always be spatially resolved, each signal, at an ever-so-slightly different frequency, was generated by an individual molecule in a slightly different local environment. This 'blinking' phenomenon, and the tiny frequency shifts observed for individual molecules, is reminiscent of single-molecule optical-spectroscopy results¹⁴.

Durkan and Welland's experiment¹, combined with previous experiments and new theory, validates this ESR–STM technique, even for scientists sceptical of the initial results of Manassen and colleagues¹¹. It points the way to exciting times ahead for spin detection and manipulation at the single quantum level.

Hari C. Manoharan is in the Department of Physics and the Geballe Laboratory for Advanced Materials, Stanford University, 476 Lomita Mall, Stanford, California 94305-4045, USA.

e-mail: manoharan@stanford.edu

1. Durkan, C. & Welland, M. E. *Appl. Phys. Lett.* **80**, 458–460 (2002).
2. Roukes, M. L. *Physics World* **14**, Feb., 25–31 (2001).
3. Bruland, K. J. *et al. Appl. Phys. Lett.* **73**, 3159–3161 (1998).
4. Stipe, B. C. *et al. Phys. Rev. Lett.* **87**, 277602 (2001).
5. Wulfskel, W. & Kirschner, J. *Appl. Phys. Lett.* **75**, 1944–1946 (1999).
6. Heinze, S. *et al. Science* **288**, 1805–1808 (2000).
7. Li, J., Schneider, W.-D., Berndt, R. & Delley, B. *Phys. Rev. Lett.* **80**, 2893–2896 (1998).
8. Madhavan, V., Chen, W., Jamneala, T., Crommie, M. F. & Wingreen, N. S. *Science* **280**, 567–569 (1998).
9. Manoharan, H. C., Lutz, C. P. & Eigler, D. M. *Nature* **403**, 512–515 (2000).
10. Yazdani, A., Jones, B. A., Lutz, C. P., Crommie, M. F. & Eigler, D. M. *Science* **275**, 1767–1770 (1997).
11. Manassen, Y., Hamers, R. J., Demuth, J. E. & Castellano, A. J. *Jr Phys. Rev. Lett.* **62**, 2531–2534 (1989).
12. Manassen, Y. *J. Magn. Reson.* **126**, 133–137 (1997).
13. Balatsky, A. V. & Martin, I. Preprint cond-mat/0112407 (2001); <http://xxx.lanl.gov>
14. Moerner, W. E. & Orrit, M. *Science* **283**, 1670–1676 (1999).

Neurobiology

A cool ion channel

Charles S. Zuker

Researchers are getting warm in their search for the details of cold detection. The discovery of a cold-sensitive ion channel will help dissect how the nervous system encodes and decodes the temperature spectrum.

The feel of a cool breeze on a hot summer day; the refreshing, soothing effects of menthol; the chill of a harsh winter night — three familiar experiences that illustrate some of the diverse sensations of 'cold'. But how do we detect chilly temperatures, and why do minty flavours have a cooling effect? Two groups now have an answer. Julius and colleagues¹ and Peier and co-workers², writing on page 52 of this issue and in *Cell*, respectively, have identified an ion channel that responds to cold stimuli and menthol. This channel is expressed in sensory neurons implicated in temperature sensation, and has the physiological and pharmacological properties expected of the mammalian cold sensor.

Cold and warm temperatures are known to be sensed by 'thermoreceptor' proteins, present on free nerve endings of the specialized (somatosensory) neurons that innervate our body surface and oral cavity. When activated, these receptors signal the sensation of heat or cold to the central nervous system. But until recently, the identity of the thermoreceptors was unclear.

The molecular biology of thermosensation has experienced a renaissance over the past few years, following the cloning and characterization of two ion channels that function as the principal heat sensors on mammalian somatosensory neurons^{3,4}. One of the channels, vanilloid receptor subtype 1 (VR1), is activated by moderate temperatures above 43 °C and by vanilloid compounds, including capsaicin, the 'hot' component of chilli peppers. The other, vanilloid-receptor-like type 1 (VRL-1), responds to temperatures above 50 °C. Together, these two receptor channels orchestrate the heat sensitivity of moderate- and high-threshold primary sensory neurons, which also function as the first stage in pain perception (reviewed in ref. 5).

By contrast, little is known about the cellular and molecular basis of cold detection. Several mechanisms have, however, been proposed to mediate the sensation of cool stimuli. These include the activation of Na⁺-selective channels known as degenerins⁶; inhibition of the (Na⁺ + K⁺)ATPase, a channel that pumps K⁺ ions into cells and Na⁺ ions out⁷; differential modulation of Na⁺ and K⁺ channels⁸; and, most recently, inhibition of K⁺ channels⁹.

Although cool temperatures do seem to affect the function of various known Na⁺

and K⁺ currents in neurons, a recent electrophysiological study¹⁰ of dissociated primary sensory neurons revealed a new, prominent inward cation current that was specifically activated by low temperatures. Notably, this current was present in all cold-responsive neurons tested, and did not belong to any of the previously proposed cold-sensing channels or pumps.

McKemy, Neuhauser and Julius¹ capitalized on this finding, and on their experience with the cloning of the VR1 heat receptor³, to search for the molecular entity that detects cold. They reasoned that if a receptor for high-temperature stimuli could be successfully isolated by an expression-cloning strategy involving capsaicin — a chemical 'mimic' of heat — then perhaps the cold receptor could be similarly isolated by using chemical compounds that elicit a cool sensation¹¹. This strategy received support when they examined cold- and menthol-activated currents in dissociated rat trigeminal sensory neurons. They found that the same subpopulation of neurons that responded to cold also responded to menthol, and that cold and menthol activated cation conductances with similar biophysical properties.

Armed with a library of complementary DNAs (copies of expressed genes) from cultured trigeminal neurons, Julius and colleagues identified a cDNA encoding a receptor that responded robustly to menthol. Lo and behold, this cDNA encoded a cation channel from the transient receptor potential (TRP) family of ion channels¹² — a familiar face to these researchers as both VR1 and VRL-1 belong to this family. More importantly, this menthol-activated ion channel, which the authors named CMR1 (cold-menthol receptor type 1), was also activated by temperatures in the range of 8–30 °C.

Humans perceive temperatures in the range of 15–30 °C as cool. In contrast, temperatures below 15 °C are classified as 'noxious' cold¹³. An obvious question is whether other TRP channels function as cold sensors, perhaps activated by noxious cold stimuli or by cooling agents such as camphor that did not stimulate CMR1. Although it is too early to know, there are many uncharacterized TRP channels encoded in mammalian genomes (reviewed in ref. 12). Indeed, it was a search for previously unknown TRP channels that led Peier *et al.*²



100 YEARS AGO

In the more or less popular accounts which have recently been given of Prof. Arrhenius' theory of cometary tails and the auroræ, it is generally stated that the smaller the diameter of the corpuscle upon which the light is falling the greater the excess of light-pressure over gravitational force. This explanation, however, holds only so long as the diameter is greater than the wave-length of light. If the diameter becomes of the same order as the wave-length, the ratio between light-pressure and gravitation follows an entirely different law. This has recently been demonstrated by Prof. Schwarzschild by an exhaustive mathematical treatment of the question... Prof. Schwarzschild's profound mathematical investigation makes it absolutely clear that the idea of minute electrically-charged corpuscles — of about one-thousandth the size of a hydrogen atom — being propelled by the sun's light towards the earth and causing the various phenomena of auroræ, Gegenschein, &c., receives no support from the mathematical point of view. But, even apart from these difficulties, it can hardly be said that the ingenious theory of Arrhenius settles the question as to the nature of the force acting on the cometary matter.

From *Nature* 6 March 1902.

50 YEARS AGO

In two previous papers, the quantum theoretical hit theory, which has been applied so successfully, for example, to gene mutations, has been applied to the problem of spontaneous cancer and cancer induced by carcinogenic hydrocarbons. The fundamental assumption of the hit hypothesis is that most, if not all, biological processes are ultimately governed by certain control centres, which are situated in the cell nuclei and are thought of as giant molecules making quantum jumps when acted upon by different agents such as chemical substances ... and ionizing radiations. In the case of mutations, these centres are now generally believed to be genes. Analogously, we have assumed in the papers referred to above that there is in each cell a certain control centre which controls the velocity of the chemical chain reactions which in turn control the velocity of growth and thus the rate of proliferation of the cells. We have now applied our theory to cancer produced by viruses, and here also the agreement with experimental findings has been very satisfactory.

From *Nature* 8 March 1952.

to identify TRPM8 (ref. 14; the same channel as that identified by Julius's group, but with a different name) as a cold receptor. Beginning with a bioinformatics approach, Peier *et al.* isolated new members of the TRP family, including TRPM8. *In situ* hybridization experiments and functional studies then

showed that this channel is a cold receptor expressed in a subset of somatosensory neurons that detect temperature and noxious stimuli.

CMR1/TRPM8 has some intriguing properties. It shows pronounced outward rectification — ions flow more readily in the

outward direction — and has relatively high permeability for Ca^{2+} ions, with little selectivity between monovalent cations (it is a non-selective cation channel). Interestingly, the channel desensitized after exposure to either cold stimuli or menthol, but the adaptation to cold temperatures could be reversed after warming to room temperature. Furthermore, cold responses were enhanced by menthol, but menthol could not stimulate the channel at warm temperatures. Thus, cold temperatures may be activating this channel by triggering progressive conformational shifts between the open and closed states, and menthol could be viewed as an allosteric modulator — it shifts the channels' activation threshold to less cold temperatures.

The discovery of a cold receptor opens the way to answering many of the outstanding questions in the field. For example, does CMR1/TRPM8 work in concert with other ion channels on the surface of temperature-sensitive neurons to shape responses to cold stimuli^{6–9}? Does the channel participate in pain pathways, and do other physiological stresses, such as the components of the 'inflammatory soup' produced by tissue injury, modulate its function? More generally, we can now start to explore how responses to warmth and cold are choreographed at the molecular level, and how the temperature spectrum is encoded and decoded by the nervous system. The next goals should be to finely map the sites of expression of the various receptors; to trace the connectivity patterns of the corresponding neurons to the spinal cord and brain; and to examine the physiological and behavioural effects of knocking out specific receptors and channels in model organisms. ■

Charles S. Zuker is at the Howard Hughes Medical Institute and Departments of Biology and Neurosciences, University of California at San Diego, La Jolla, California 92093-0649, USA.

e-mail: charles@flyeye.ucsd.edu

- McKemy, D. D., Neuhauser, W. M. & Julius, D. *Nature* **416**, 52–58 (2002); online 10 February 2002 (DOI 10.1038/nature719).
- Peier, A. *et al.* *Cell* online 11 February 2002 (DOI 10.1016/S0092867402006529).
- Caterina, M. J. *et al.* *Nature* **389**, 816–824 (1997).
- Caterina, M. J. *et al.* *Nature* **398**, 436–441 (1999).
- Julius, D. & Basbaum, A. I. *Nature* **413**, 203–210 (2001).
- Askwith, C. C. *et al.* *Proc. Natl Acad. Sci. USA* **98**, 6459–6463 (2001).
- Pierau, F. K., Torrey, P. & Carpenter, D. O. *Brain Res.* **73**, 156–160 (1974).
- Carpenter, D. O. *Fed. Proc.* **40**, 2808–2813 (1981).
- Viana, F., de la Peña, E. & Belmonte, C. *Nature Neurosci.* **5**, 254–260 (2002); online 11 February 2002 (DOI 10.1038/nn809).
- Reid, G. & Flonta, M. L. *Nature* **413**, 480 (2001).
- Hemsel, H. & Zotterman, Y. *Acta Physiol. Scand.* **24**, 27–34 (1951).
- Clapham, D. E., Runnels, L. W. & Strubing, C. *Nature Rev. Neurosci.* **2**, 387–396 (2001).
- Chen, C. C., Rainville, P. & Bushnell, M. C. *Pain* **68**, 33–43 (1996).
- Tsavalier, L. *et al.* *Cancer Res.* **61**, 3760–3769 (2001).

Biogeochemistry

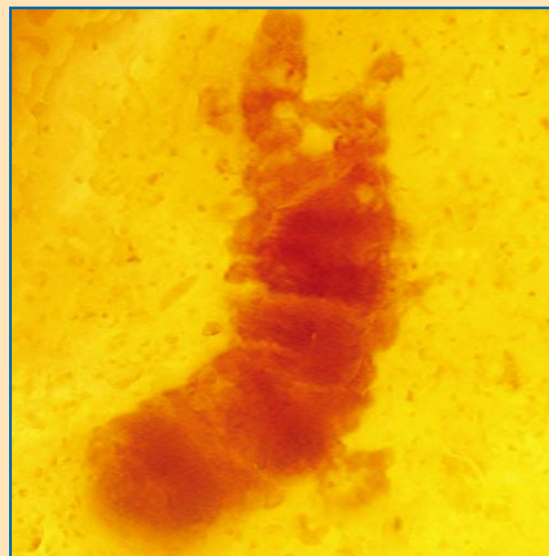
That's life?

This could be a picture of one of the earliest-known fossils — a microbe, 60 μm long, that is almost 3,500 million years old. On the other hand, it could be just a flaw in the rock. It all depends on whom you talk to, as epitomized by two papers in this issue.

The more optimistic view is taken by William Schopf and colleagues (*Nature* **416**, 73–76; 2002). In the early 1990s, Schopf caused a sensation with reports of a diverse bacterial flora from the 3,465-million-year-old Apex cherts of Western Australia. At the time, all he had to go on was morphology. This was always controversial, given that bacteria have little morphology to begin with. As a consequence, it is hard to tell the difference between a bacterium — especially a fossil bacterium — and a bubble. This is why Schopf has devised authenticity criteria that are based on bacterial habit. If you have one bacterium, for example, you will usually have hundreds: isolated blobs purporting to be bacteria usually turn out to be artefacts.

Since then, Schopf and colleagues have sought to back up the morphology using laser-Raman imaging, a technique in which the chemical composition, as well as the structure of the fossils, can be mapped in two dimensions. After several tests of the technique on less controversial fossils, Schopf *et al.* have used the method on the Apex chert material. They find that the fossils have the composition to be expected if they were made of organically derived carbon.

This will not be the end of



the controversy, however, as demonstrated by the report from Martin Brasier and colleagues (*Nature* **416**, 76–81; 2002). Using the original Apex chert material, as well as newly collected specimens, they find that the rocks in which the fossils were found come from a vein that may have been produced hydrothermally, that is, by the action of heated water on minerals. The carbon-isotope signature is consistent with a biogenic origin, one derived from living organisms. But other studies suggest that this was an environment in which carbon dioxide produced by volcanic action was transformed into isotopically light carbon at 250–350 °C. The bottom line is that although the carbon looks organic, it need not be. Similarly, Raman spectroscopy shows that although the material in the fossils could be biogenic, it could equally well be amorphous graphite.

Brasier *et al.* even dismiss the morphology. They suggest

that the fossils, as a whole, have a random orientation that is not characteristic of bacterial behaviour, in which the cells tend to line up in one direction or another; the individual cells have a range of strange, even branched, morphologies; and what look like filaments with internal divisions may be the result of interleaving quartz and graphite sheets. Many authors agree that there is isotopic evidence for biogenic activity in the Archaean (that interval of Earth history before 2,500 million years ago). But Brasier and colleagues, at least, say that the Apex chert fossils aren't really fossils at all.

Given that Schopf was one of the first to cast doubt on the biogenicity of another celebrated suite of purported microfossils — in the martian meteorite ALH84001 — it is ironic that his own work should be subjected to such scepticism. But that is the name of the game for claims of life at the extremes of time and space. **Henry Gee**

Astronomy

How big stars are made

Susana Lizano

We know little about the formation of stars that are many times heavier than our Sun. A new model, which estimates how long massive stars take to form, could be tested with data from telescope arrays.

The formation of low-mass stars such as the Sun is relatively well understood¹, because there are several regions near the Earth in which such stars have been observed in detail. But for massive stars (more than eight solar masses), which are generally formed at greater distances from our planet, the challenge to understand their formation process remains. On page 59 of this issue, McKee and Tan² present a dynamic model for the creation of massive stars, and deduce the timescale on which such stars form. Their work is not only relevant for star formation, but also has a direct impact on other fields of astrophysics, such as the formation, evolution and chemical enrichment of galaxies.

Massive stars are found in small groups called OB associations, together with hundreds of low-mass stars (Fig. 1). The closest and best-known example is the Trapezium cluster in the Orion nebula, a star-forming molecular cloud 1,500 light years from Earth. High-resolution observations over the past decade have revealed extreme physical conditions in these regions.

For an OB association to form³, it seems that thousands of solar masses of gas (mainly molecular hydrogen) must be compressed in a region smaller than 1 parsec (1 pc = 3.08×10^{18} cm). Because young, massive stars produce a large fraction of their luminosity at frequencies that are high enough to ionize and heat the surrounding gas cloud, one of the first signs that massive stars are forming is the appearance of compact ionized regions. Fortunately, these can easily be detected at centimetre wavelengths, where the gas clouds are transparent. Hot molecular cores smaller than 0.1 pc that contain hundreds of solar masses of gas have been discovered near the compact ionized regions⁴, and these are thought to indicate an even younger phase, in which massive stars are still growing by accretion from infalling material.

There are many questions that the star-formation community is striving to answer: are massive stars just larger versions of low-mass stars, with the same, but scaled-up, features, such as the protoplanetary disks, bipolar jets and outflows observed in their lighter relations? Do high- and low-mass stars form simultaneously in OB associations? Do the violent winds and powerful ionizing radiation produced by young massive stars prevent further star formation

in the cloud, or do shocks travelling through the cloud induce the formation of new generations of stars? And what determines the fraction of gas in the cloud that will eventually be transformed into stars? The answers to these questions are not only fundamental to our understanding of star formation but also have wide implications throughout astrophysics. Understanding how gas is transformed into stars and how these new stars interact with the gas in their galaxy — perturbing, dispersing and enriching it with heavy elements — are key issues.

McKee and Tan² propose that stars larger than eight solar masses are formed within small dense cores of molecular gas that are initially in hydrostatic equilibrium (when gravitational forces are exactly balanced by the counteracting gas pressure and magnetic forces). Such protostellar cores are found inside high-pressure regions in molecular clouds that lie in the plane of our Galaxy.

The high ambient pressures imply that the hydrostatic molecular cores are very dense, tens of thousands of times denser than the parent clouds. When such a core becomes gravitationally unstable, it collapses very quickly, and a protostar forms at the centre of the core. An intense accretion phase begins, as core material falls onto the protostar. The protostar accretes huge amounts of material at rates that increase with time, reaching values 100 to 1,000 times larger than those of low-mass stars. High accretion rates in the formation of massive stars have been inferred from observations and other theoretical models, but now McKee and Tan provide a sound physical explanation for such rates. The high accretion rates allow the star to grow before the pressure of its own radiation exerted on the infalling dust is able to push back and reverse the accretion. The authors show that stars of 10 to 30 solar masses can form easily and rapidly inside these cores, on timescales of the order of 100,000 years.

In their model, McKee and Tan assume that the final mass of a star is determined by the initial core mass, with half of the core mass going into the star. The remaining core mass is dispersed by a powerful stellar wind, which is 100 to 1,000 times stronger than that generated by low-mass protostars. But an alternative model assumes that the core contains hundreds of solar masses of gas, many times more than the final mass of the star forming at the centre; then it is the powerful

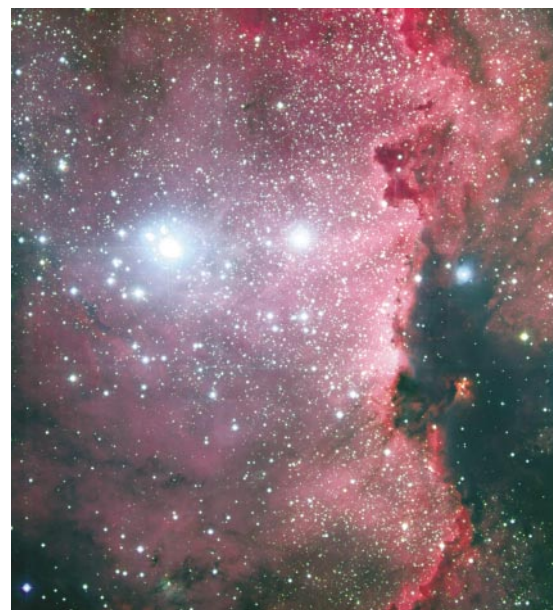


Figure 1 In the southern Milky Way, 4,000 light years from Earth, the constellation of Ara (the Altar) is a cradle of newborn stars. The blue stars on the left in this image are young and massive. Intense radiation from these stars is ionizing the surrounding cloud of hydrogen gas, producing the characteristic red glow. McKee and Tan² have calculated that such stars take 100,000 years to form.

stellar wind that prevents the infall of all the excess mass onto the central star⁵.

The question remains as to what physical process produces these extremely dense hydrostatic cores. Is it a gradual condensation process, as in the case of low-mass stars, where the collapse is delayed by magnetic or turbulent pressure until very high densities are reached? Or is it a fast process caused by shocks within a turbulent medium? In the latter case, it has been argued that the cores might not be hydrostatic⁶.

In the meantime, it is crucial to proceed with sensitive observations, at high spatial resolution, and compare them with theoretical models of massive-star formation. One useful clue is the energy flux emitted by warm dust in the cores, which can be measured in ever greater detail as instrumentation improves. An important test will be to compare the emission predicted by the McKee and Tan model with dust-emission profiles from hot cores. High-resolution data from existing millimetre-scale interferometers⁷ are accumulating, but the final answer may have to wait for the new generation of millimetre and submillimetre interferometers, such as the Submillimetre Array, now taking data in Hawaii, and the Atacama Large Millimetre Array, currently under construction in Chile. ■

Susana Lizano is at the Instituto de Astronomía UNAM, Campus Morelia, Morelia, Michoacán 58190, Mexico.
e-mail: s.lizano@astrosmo.unam.mx

1. Lada, C. & Shu, F. *Science* **248**, 564–572 (1990).
2. McKee, C. F. & Tan, J. C. *Nature* **416**, 59–61 (2002).
3. Plume, R., Jaffe, D. T., Evans, N. J., Martín-Pintado, J. & Gómez-González, J. *Astrophys. J.* **476**, 730–749 (1997).
4. Kurtz, S., Cesaroni, R., Churchwell, E., Hofner, P. & Walmsley, C. M. in *Hot Cores and the Earliest Phases of High-Mass Star Formation, Protostars and Planets IV* (eds
- Mannings, V., Boss, A. P. & Russell, S. S.) 299–326 (Univ. Arizona Press, Tucson, 2000).
5. Osorio, M., Lizano, S. & D'Alessio, P. *Astrophys. J.* **525**, 808–820 (1999).
6. Ballesteros-Paredes, J., Vázquez-Semadeni, E. & Scalo, J. *Astrophys. J.* **515**, 286–303 (1999).
7. Molinari, S. *et al.* *Astrophys. J.* **505**, 39–42 (1998).

Cell biology

Ripping up the nuclear envelope

Jennifer Lippincott-Schwartz

During cell division, the membranes that surround the nucleus must be dismantled to allow the DNA housed inside the nucleus to be partitioned into two daughter cells. New work shows how this happens.

One of the many events that take place as cells divide is the breakdown of the nuclear envelope — the membrane that surrounds the nucleus. Given the elaborate structural organization of the nuclear envelope¹, this is not a simple task. Rather unusually, the envelope consists of not one but two layers of membrane, which are punctuated by pores through which molecules can be moved into and out of the nucleus. Not only that, but branching off the surface of the outer layer is another membranous compartment, the endoplasmic reticulum. So how do cells dismantle the nuclear envelope?

Beaudouin and colleagues² and Salina and co-workers³ have put forward a likely model in a recent issue of *Cell*. The authors provide evidence that a molecular motor protein, cytoplasmic dynein, is used to tear the envelope, transporting pieces of membrane away along filamentous structures called microtubules. The results reveal a new type of function for a molecular motor, and may provide insight into how other organelles are taken apart.

But why do dividing cells need to break down the nuclear envelope at all? The answer lies in the mechanics of nuclear division (mitosis). Before mitosis begins, the cell duplicates its chromosomes. During mitosis, the two chromosome copies ('sister chromatids') are partitioned to opposite poles of the cell, which then divides into two. To achieve such partitioning, the sister chromatids are first attached to the spindle — a structure composed of microtubules — that itself links to two microtubule-organizing centres, or 'centrosomes'. The microtubules then shrink back towards the centrosomes, pulling the sister chromatids apart. The problem is that, in many cells, the centrosomes are located outside the nucleus (that is, in the cytoplasm) during early stages of mitosis, whereas the sister chromatids are, of course, inside the nucleus. Unless the nuclear envelope breaks down, microtubules cannot attach to and partition the chromatids.

The mechanism by which the nuclear

envelope breaks down has been the subject of much debate⁴. The traditional view is one involving vesiculation, in which, at the start of mitosis, nuclear-envelope membranes are converted into small vesicles that disperse into the cytoplasm. The idea has received support from subcellular 'fractionation' studies. But direct visualization of mitotic cells expressing fluorescent envelope proteins has shown that these proteins move freely throughout a single, continuous endoplasmic reticulum (ER) membrane system, with no evidence of vesicular intermediates^{5,6}. These results have led to an alternative model that involves release and diffusion of envelope proteins and lipids into the ER, probably as a result of phosphorylation events, which abolish protein interactions essential for nuclear-envelope integrity.

But there is a problem with this model, which has to do with the need to quickly remove the nuclear envelope from the vicinity of chromosomes to allow the spindle apparatus to assemble and function. Diffusion into the ER takes time and does not immediately remove the membranes around the nucleus. Beaudouin *et al.*² and Salina *et al.*³ provide a neat solution to the problem. Together they show that, after initiating nuclear-envelope breakdown by a force-driven tearing process, dynein quickly transports pieces of envelope, still attached to the ER, away from chromosomes along microtubules. At the same time, envelope lipids and proteins are mixed with ER components.

The first hint of a mechanistic link between microtubules and nuclear-envelope breakdown came nearly 40 years ago, when classic electron-microscopic studies showed an intimate connection between the spindle and nuclear envelope, as well as the presence of deep invaginations in the nuclear envelope near centrosomes⁷. It was first thought that these invaginations represent sites at which spindle microtubules 'pierce' the membrane⁸. But, when Beaudouin *et al.* looked at nuclear-envelope breakdown in living cells, they discovered something quite

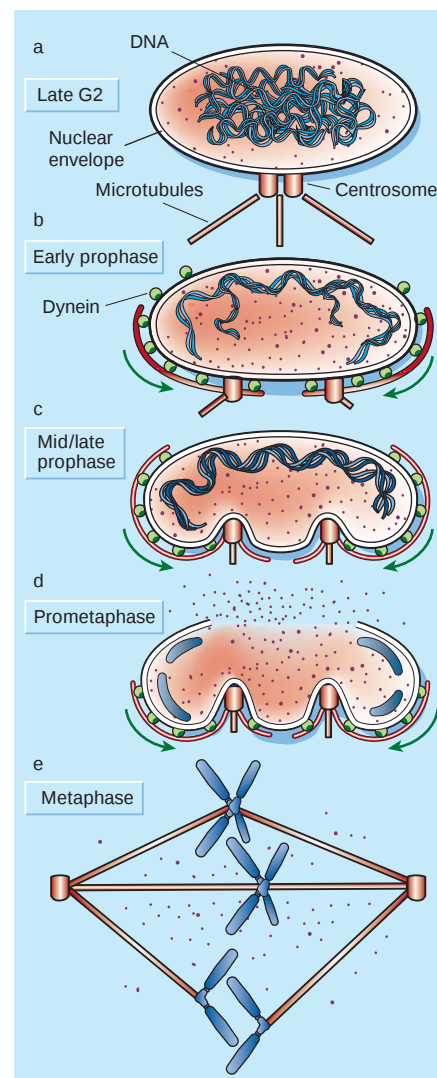


Figure 1 How cells dismantle the nuclear envelope during the phases (shown in boxes) of cell division, as proposed by Beaudouin *et al.*² and Salina *et al.*³. a, Chromosomes have been duplicated and are in an 'uncondensed' state. The centrosome has likewise been duplicated and is found just outside the nucleus. b, The molecular motor dynein is recruited to the nuclear envelope and interacts with microtubules. c, Dynein pulls nuclear-envelope components along the microtubules towards the centrosomes, causing invaginations. d, Withdrawal of membrane from the opposite side of the nucleus creates membrane tension, causing tearing. The nuclear envelope opens up catastrophically as the membranes continue to be pulled towards centrosomes. From early prophase onwards, the chromosomes are condensing into well-resolved forms and, after nuclear-envelope breakdown, can be attached to microtubules and segregated (e).

different (Fig. 1): invaginations result when spindle microtubules pull on nuclear membranes in the direction of centrosomes. More precisely, the authors found that microtubule-dependent forces mechanically stretch the nuclear lamina — the tight mesh

of cytoskeletal proteins just inside the nucleus — and deform nuclear membranes, perturbing the distribution of nuclear-pore complexes and the ER.

Subsequent perforation of the nuclear envelope occurs not at the sites of invagination but on the opposite side of the nucleus, where tension is greatest. Once torn in this way the nuclear envelope catastrophically loses its shape, and cytoplasmic proteins flood into the nuclear space. It is not clear exactly how the initial perforation comes about. One clue comes from the finding that, in starfish embryos, the envelope becomes permeable to macromolecules (which are normally excluded from the nucleus by nuclear-pore complexes) before it is perforated⁶. So it is possible that microtubule-dependent changes in envelope structure might induce localized disassembly of nuclear pores, creating an epicentre for tearing.

The influx of cytoplasmic molecules that occurs after perforation might facilitate several mitotic processes, including chromosome condensation and spindle formation. Indeed, Beaudouin *et al.* show that chromosome condensation, which begins before nuclear-envelope breakdown, accelerates threefold after perforation. All in all, microtubule-dependent tearing seems to allow cells to tightly coordinate nuclear-envelope breakdown with spindle formation and chromosome dynamics.

Meanwhile Salina *et al.*³ looked at the molecular basis of microtubule-dependent nuclear-envelope tearing, and find that cytoplasmic dynein becomes redistributed to the cytoplasmic face of the envelope before it invaginates. Once there, dynein associates with dynactin — an activator of dynein-mediated transport processes⁹ — and pulls nuclear membranes and other envelope components along microtubules towards the centrosome.

It could prove challenging to work out exactly how dynein binds to the nuclear envelope. This protein is involved in many cellular processes and interacts with many structures and proteins. One possibility is that it associates with membranes through spectrin, a protein that — with dynactin — forms a lattice around organelles¹⁰. Whatever the case, dynein clearly transmits force across the nuclear envelope, as tearing disrupts both the inner and outer layers. This points to the possibility that dynein interacts with molecules associated with the nuclear-pore complexes, which span both layers. Another issue that needs to be resolved is how dynein is redistributed to the nuclear envelope from its usual cytoplasmic location in a cell-division-dependent way. Such redistribution might be controlled, at least in part, by the phosphorylation of dynein¹¹.

Might the mechanism described by Beaudouin *et al.*² and Salina *et al.*³ be more generally applicable? The ER, mitochondria

and Golgi apparatus all need to partition into daughter cells during mitosis, and a pulling and tearing mechanism dependent on dynein and microtubules could be at work here, too. In fact, the Golgi apparatus does split into two populations that follow centrosomes to different positions in the cell before being absorbed into the ER¹², and mitotic Golgi remnants have been shown to be swept towards centrosomes¹³. It remains to be seen whether or not dynein is involved. But its function in nuclear-envelope breakdown is now clear, and provides a good illustration of how two fundamental cellular processes — motor-driven transport and organelle breakdown — can be coordinated. ■

Jennifer Lippincott-Schwartz is in the Cell Biology

and Metabolism Branch, NICHD, National Institutes of Health, 18 Library Drive, Bethesda, Maryland 20892-5430, USA.
e-mail: jlippin@helix.nih.gov

1. Gerace, L. & Burke, B. *Annu. Rev. Cell Biol.* **4**, 335–374 (1988).
2. Beaudouin, J., Gerlich, D., Daigle, N., Eils, R. & Ellenberg, J. *Cell* **108**, 83–96 (2002).
3. Salina, D. *et al.* *Cell* **108**, 97–107 (2002).
4. Collas, I. & Courvalin, J. C. *Trends Cell Biol.* **10**, 5–8 (2000).
5. Ellenberg, J. *et al.* *J. Cell Biol.* **138**, 1193–1206 (1997).
6. Terasaki, M. *et al.* *Mol. Biol. Cell* **12**, 503–510 (2001).
7. Robbins, E. & Gonatas, N. K. *J. Cell Biol.* **21**, 429–463 (1964).
8. Georgatos, S. D., Pyrpasopoulou, D. & Theodoropoulos, P. A. *J. Cell Sci.* **110**, 2129–2140 (1997).
9. Karki, S. & Holzbaur, E. L. *Curr. Opin. Cell Biol.* **11**, 45–53 (1999).
10. DeMatteis, M. A. & Morrow, J. S. *J. Cell Sci.* **113**, 2331–2343 (2000).
11. Addinall, S. G. *et al.* *J. Biol. Chem.* **276**, 15939–15944 (2001).
12. Zaal, K. J. M. *et al.* *Cell* **99**, 589–601 (1999).
13. Shima, D. T. *et al.* *J. Cell Biol.* **141**, 955–966 (1998).

Human evolution

Tangled genetic routes

Rebecca L. Cann

It is generally accepted that early human evolution took place in Africa, with human populations spreading from there. Using genetics to trace events in more detail remains a challenging task.

How best can the riches of human genetics be mined for information about our history and geography? In the tradition of human population genetics advanced over the past 50 years¹, the search is on for evidence of the isolation, adaptation and dispersal of populations over time. In this case, there is no digging for fossils, no three-dimensional reconstructions of skulls and such, just the requirement to collect 10-ml samples of blood. However, we are sometimes at a loss to disentangle population structure from population history. The diversity of DNA sequences in a modern population is an accumulation of events in the remote past. But it is unclear that the same forces that change allele frequencies (the incidence with which different gene forms occur) also change the genealogies of genes.

On page 45 of this issue², human phylogeographers — who study the geographical distribution of genealogical lineages using DNA sequence variation³ — acquire a new approach to a recurrent problem. Alan Templeton² has analysed 11 different human gene trees with the program GEODIS⁴. His aim has been to assess the strength of the geographical signals they contain, using tests with nested clades (that is, groups of haplotypes — linked alleles — that are arranged by successively increasing numbers of mutations). His analysis provides strong genetic support for describing the geographical centre of our species as African, with at least two major population expansions from that continent about 600,000 and 95,000 years ago. He also establishes a platform to make specific estimates for important parameters of

population structure, including the level of gene flow between populations isolated by distance. As a strong supporter of the idea of an African origin for modern people, in the words of Hamlet, “I eat the air, promise-cramped”.

Current limitations in the methods used to reliably extract ancient DNA⁵ have led molecular geneticists to concede the direct study of the earliest stages of our lineage's history to palaeoanthropology. So events well before 2 million years ago remain the province of those who hunt for and interpret fossil remains. But there is also a contentious debate over modern human origins that centres on the time period from 1.7 million to 20,000 years ago and the emergence of anatomically modern people. Genetic diversity in the human population today is consistent with a model of expansion, predominantly from Africa, between 100,000 and 50,000 years ago. From this breeding population of as few as 10,000 individuals, there followed a second expansion in Europe about 21,000 years ago^{6,7}.

Oscillations in climate are assumed to have resulted in an increased isolation of different groups, which would have promoted the ‘fixation’ of local adaptations in morphology, physiology and behaviour. Geographically distinct populations in Europe, Africa, South Asia, China and Australia could have had separate evolutionary trajectories if there were not enough migrants in each generation to spread any new mutations to populations in other regions. Proponents of multiregionalism believe that, after the initial expansion from Africa, populations continued to evolve

in different parts of the world, including Africa. This view places special emphasis on the role of migration in transferring new mutations affecting important human genes arising in geographically isolated groups. More migrants came from Africa, where there were bigger populations, in which greater numbers of new mutations would arise. But do the inferences of differential migration that are essential to this view hold up to scrutiny?

Templeton and other proponents of multiregionalism have criticized Afrocentric interpretations as incomplete, and an inaccurate reflection of how modern populations are and have been structured; they claim that researchers are confusing history with geography. This problem is one of the reasons that an approach incorporating methods known as 'spatial autocorrelation' and 'multidimensional scaling' has gained popularity. Here, patterns of genetic variation are compared with those predicted from theoretical models, and then used to estimate parameters such as levels of inbreeding, natural selection and gene flow⁸. But this approach ignores certain information in the sequences themselves. So it has not been as useful as phylogenetic analysis for identifying a major event in the evolution of allelic lineages within a species — such as a large population 'bottleneck' (in which, at some point in time, comparatively few individuals of a species exist, and so genetic diversity is limited).

In human evolutionary studies, methods that are commonly used to identify differences between species are often extended to the study of variation within a species. Templeton sees this as a major shortcoming in the genetic approach to human evolution. He argues that some morphological traits would fail to show regional continuity between populations over time, if they are not under the strong influence of natural selection, and that gene flow from dispersing and expanding groups is balanced by the presence of random genetic drift (the random loss of alleles over time through small population size). In his view, an estimate⁹ that only 90% of the haplotype trees in the nuclear genome demonstrate African roots is clear evidence that regionally isolated, archaic populations were not completely replaced by immigrants of ultimate African origin. Templeton attempts to break this impasse in understanding by showing that what is needed is an analytical tool favouring neither model on an *a priori* basis.

What assumptions lie behind his GEO-DIS analysis? First, a null hypothesis states that there would be no expected association between geography and a tree. Second, a plausible set of trees must be produced. Third, it is essential that the geographical sampling design is adequate. Also, the power of the analysis should increase if sampling includes greater numbers of unlinked regions of the genome — that is, regions inherited independently of others during

cell division. Do these assumptions match the application in this analysis?

Gene positions, or loci, on chromosomes 14, 16, 21, X and Y, and those in mitochondrial DNA, were chosen, representing a maximum scan of about 13% of the human genome. Sequences that recombine (those essential to pigmentation, the immune response, oxygen consumption and glycolysis) were considered equally with those that do not (maternal and paternal markers), and with the apparently non-functional pieces of DNA known as microsatellite markers that are inherited from both parents and are presumed to be neutral. Some of the analysis required the pooling of samples from individuals from different locations in order to perform the nested cladistic analysis, and sample size varied from 35 to 1,544 individuals.

It therefore comes as no surprise that evidence for range expansion, long-distance dispersal and isolation by distance varied in strength between the genetic loci sampled. In this respect, Templeton's analysis shows how our current understanding, based on loci that are non-concordant in pattern and process even in the same populations, limits certainty about a global reconstruction of human evolution.

Any tool that helps to clarify the magnitude of genetic drift, population migration or natural selection, and directs the researcher to what is needed (more donors, more sequence), is helpful. But perhaps Templeton was over-ambitious in the scale of his analysis. Most specialists would consider the reconstruction of the human population expansion into Polynesia to be a relatively simple task compared to what Templeton has attempted. But I know only too well that analysis of blood from 35 donors taken from just one island was insufficient to describe the major roots of diversity on that island, although the Polynesian expansion was quite recent. Perhaps we will need a demonstration that GEODIS reveals the composite picture agreed upon by archaeology, genetics and linguistics in this area of the world before we can settle on how to interpret the varied signals uncovered by Templeton's analysis on a global scale. ■

Rebecca L. Cann is in the Department of Genetics and Molecular Biology, Biomedical Sciences Building A-204, University of Hawaii, Manoa, 1960 East-West Road, Honolulu, Hawaii 96822, USA. e-mail: rcann@hawaii.edu

1. Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. *The History and Geography of Human Genes* (Princeton Univ. Press, Princeton, New Jersey, 1994).
2. Templeton, A. R. *Nature* **416**, 45–51 (2002).
3. Avise, J. C. et al. *Annu. Rev. Ecol. Syst.* **18**, 489–522 (1987).
4. Posada, D., Crandall, K. A. & Templeton, A. R. *Mol. Ecol.* **9**, 487–488 (2000).
5. Cooper, A. & Poinar, H. *Science* **289**, 1139 (2000).
6. Harpending, H. et al. *Proc. Natl Acad. Sci. USA* **95**, 1961–1967 (1998).
7. Jorde, L., Bamshad, M. & Rogers, A. *BioEssays* **20**, 126–136 (1998).
8. Bertorelle, G. & Barbujani, G. *Genetics* **140**, 811–819 (1995).
9. Takahata, N., Lee, S.-H. & Satta, Y. *Mol. Biol. Evol.* **18**, 172–183 (2001).

Daedalus

Solid cooling

The world's domestic refrigerators and air-conditioners all work on the same principle. A liquid is evaporated under low pressure, absorbing the latent heat of evaporation. Elsewhere the vapour is condensed under high pressure, returning that latent heat to a different system. The cycle then repeats. The snag is that the best fluid to use is a halocarbon, which damages the ozone layer. This wouldn't matter except that all refrigerators — especially those in cars — leak. New halocarbon is needed to top them up. So the whole halocarbon problem is merely one of bad plumbing.

Daedalus now has a way out: invent a solid working medium, which cannot leak, and the problem will go away. Most solids are so incompressible that only a small change of temperature with pressure seems feasible. Rubber, of course, warms when you stretch it, as the entropy of the molecules is increased. But Daedalus recalls that carbon nanotube is a valuable probe in an atomic-force microscope. If the force is excessive, the nanotube buckles and folds up totally. This solid collapse destroys the chemically resonant structure of the tube, but is fully reversible — remove the load and it springs straight and true again. Its resonance energy must be absorbed as heat in its collapse, and re-emitted in its recovery.

So DREADCO chemists are packing as much carbon nanotube as possible into a flexible solid, possibly rubber, looking for truly dramatic changes in temperature when the composition is deformed. A small bending or squashing should collapse the nanotubes strongly, so they lose their resonant energy and absorb heat. When the solid relaxes, the nanotubes will reform, lose energy as they reorganize, and deliver it as heat again. The result should be a working solid for a refrigerator.

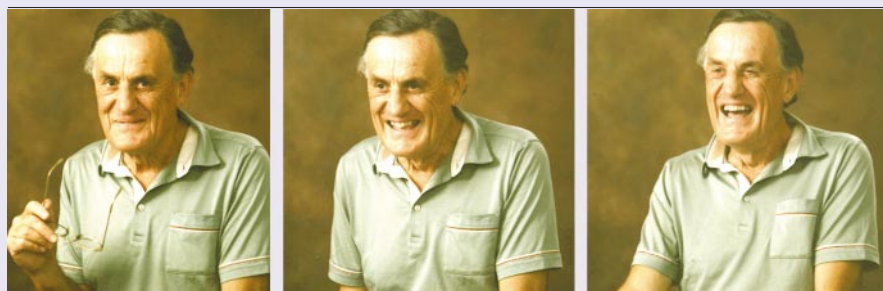
A solid-state refrigerator poses unusual design problems. A big disc, cylinder or belt will carry the working solid, and will be rotated by a motor. Wheels on the periphery will deform it where cooling is desired, and release it to recover heat. The whole thing will be simple rather than efficient.

The DREADCO team reckon they will be lucky to get more than a few tens of watts of cooling; but with no plumbing to worry about, the new refrigerator should find a useful niche in the market. The automotive market is the obvious one. Efficiency is not very important, the device will not leak or stop working, and cooling power can be increased merely by speeding it up.

David Jones

Obituary

Robert Hanbury Brown (1916–2002)



Pioneer in radar and observational astronomy

Robert Hanbury Brown, a giant from a golden period of innovation in astronomy, died on 16 January this year, at the age of 85.

Hanbury, as he was always known, was born in India in 1916, the son of an army officer. Forsaking his schoolboy plans to become a classics scholar, he graduated in engineering from the University of London in 1935. He then worked on the secret development of the coastal radar — Chain Home — which was to prove vital in the 1940 Battle of Britain. By then, Hanbury was working with a group developing a shorter-wavelength radar that could be installed in aircraft. His splendid autobiography, *Boffin* (Adam Hilger, 1991), gives a vivid account of the trials and triumphs of this work, which by 1941 gave night fighters of the Royal Air Force the edge over the German bombers.

Hanbury's contributions also included work on the polarization of radio waves, crucial for the optimal design of aerials on equipment for air-to-surface surveillance and in the detection of ships and submarines. During his secondment (1942–47) to the US Naval Research Laboratory in Washington DC, his work led directly to the NATO 'identification friend or foe' (IFF) system of today, along with the civilian system of ground-based air-traffic control now used throughout the world.

But all this was a prelude. Returning to Britain, Hanbury sought to pursue a higher degree in university research. His work having been secret, however, he had no publications. His enquiries brought him together with one of us (B.L.), who had returned to the University of Manchester with military radar equipment and built a 218-ft radio telescope at Jodrell Bank. Finding ways to minimize the background noise that bedevilled early efforts, Hanbury and his student Cyril Hazard showed that cosmic radio waves were emanating from the Andromeda spiral galaxy, 2 million light years away and far outside our own Galaxy — which,

until then, had been thought to be the origin of all such emissions.

This discovery, using a fixed paraboloid dish, was the happy retort to those questioning the feasibility of the huge steerable radio telescope B.L. was then evangelizing for. Following years of struggle, this pioneering 250-ft, steerable dish was completed in 1957. Hanbury was among those to use it. In particular, he and Hazard played a significant part in the discovery of quasars — 'quasi-stellar objects', now taken to be the most distant, and so oldest, observable objects in the Universe.

Early on, Hanbury had also begun to think about a radio interferometer, which could measure the angular size of the sky's two strongest radio sources, Cygnus A and Cassiopeia A. For all anyone knew, this might have required an interferometer baseline of a thousand kilometres or more. The technical difficulty lay in combining the radio signals, received at two widely separated points, with the required phase stability. Hanbury had the idea of measuring the correlation of the fluctuations in intensity, a process that does not require the phase information needed in approaches based on detecting signal amplitudes. This deceptively simple solution raises profound questions, and Hanbury collaborated with the theoretical physicist Richard Twiss to put it all on a sound basis. Surprisingly quickly, by 1952 the intensity interferometer was built, and had shown the radio sources to be so extended that baselines of only a few kilometres were needed (so that earlier methods would have been adequate). In Hanbury's words, he had built a sledgehammer to crack a nut.

What came next ensured that Hanbury's name will always be in astronomy texts. He and Twiss applied the same ideas to optical astronomy, but immediately encountered objections that their proposals violated basic physical laws. The experimental approach

depended on correlating light quanta, or photons. In essence, it required photons to arrive in pairs: astronomers thought they might; some physicists said they could not. At the same time as the Hanbury Brown–Twiss equation demonstrated the feasibility of pair-arrival, Hanbury's wife Heather symbolically produced twin sons.

In a major astronomical 'first', Hanbury demonstrated the feasibility of his intensity interferometer by measuring the diameter of the 'dog star', Sirius, under the unusually favourable sky conditions of the harsh British winter of 1955–56. He then sought more interesting stars in the Southern Hemisphere, under less cloudy skies. So he left his personal chair at Manchester, and moved to the University of Sydney in 1962, joining the extraordinary physics department being put together there by Harry Messel. This was the first really multi-professorial department in Australia (one of us — R.M. — being there at this time as a young faculty member).

Hanbury's next interferometer was built in a sheep paddock, outside Narrabri, New South Wales. Each of its two telescopes was 23 ft in diameter and composed of 250 mirrors, and they moved on a track 600 ft in diameter. The local wildlife of the outback provided both diversion and disruption. And drought and flood were a feature of the area. Nonetheless, the interferometer was a winner, with the diameter of tens of stars being measured, and a new scale of stellar temperatures being established.

During these years, Hanbury and his family shuttled between Narrabri and his home at Forty Baskets Beach, with its splendid view of Sydney harbour. Much of the shuttling was done in his beloved Alvis 1952 drophead coupé: given his driving style, the ride down the potentially lethal 1-in-4 drive to this house was unforgettable. Hanbury loved the warmth of Australia, and it reciprocated. He was an early recipient of its highest honour, Companion of the Order of Australia (the order is a cheerfully confused manifestation of Oz republicanism from the 1970s, the Queen being its head).

Five years after his retirement in 1981, he and his wife returned to Britain, and lived quietly in Hampshire. Hanbury was a superb and imaginative engineer, a natural astronomer, and a true visionary. He was also, as we and so many others can attest, a really lovely man.

Bernard Lovell and Robert M. May
Bernard Lovell is at *The Quinta*, Swettenham, Congleton, Cheshire CW12 2LD, UK.

Robert M. May is in the Department of Zoology, University of Oxford, Oxford OX1 3PS, UK.
e-mail: robert.may@zoo.ox.ac.uk

Searching for new islands in sea ice

Coastlines concealed in polar seas are now more accessible to cartography.

Tobias Island, discovered in 1993 by the German research vessel *RV Polarstern*, is a system of low-lying banks and shoals hidden in sea ice 70 km off the northeastern coast of Greenland. Here we use satellite radar interferometry and airborne laser scanning to show that this island is 2 km long and 35 m high — much larger than was originally reported¹. We have also been able to pinpoint the exact location of a stable area where a new group of small islands may be hidden. This demonstrates that satellite radar interferometry is an effective tool for finding ice-covered islands as well as for mapping them.

Tobias Island was only recently discovered because it is remote and is permanently surrounded by sea ice, and conditions there are frequently foggy and cloudy. Satellite radar is ideal for the investigation of such areas as it is independent of daylight and is able to 'see' through clouds. Satellite radar interferometry relies on the comparison of two images², allowing changes in the propagation path length from the satellite to the terrain to be detected with an accuracy of a fraction of a wavelength. The radar echo in one pixel is the sum of the echoes from many small scatterers. The terrain thus needs to be stable to ensure that a phase change in the reflected radar signal is caused by different propagation delays, and not simply by a change or rearrangement in the radar echoes from all of the small scatterers. For the European Remote Sensing Satellite 1/2 (ERS-1/2) radar data used here, the wavelength is 5.7 cm and images with a time separation of 1 or 3 days are combined.

The most obvious way to find new islands is to search for areas that have high interferometric radar correlation. This would indicate bare rocks and, in periods without melt events, snow-covered islands, stranded sea-ice floes and icebergs, and shore-fast sea ice. Drifting sea ice and open-water areas do not generate any correlation. It is more difficult to discriminate between shore-fast ice, stranded ice and land, all of which potentially reveal a high correlation.

We make use of the constant phase given by land and stranded ice, as opposed to the varying phase found in images from shore-fast sea ice that has been displaced by the tide and deformed by wind stress. A low-correlation flexure zone separates stranded and floating ice; a time series is required to differentiate land from stranded ice. Correlation and phase images can therefore be used to detect island candidates and to define the maximum extent of their coastlines.

Our analysis of radar data from an area

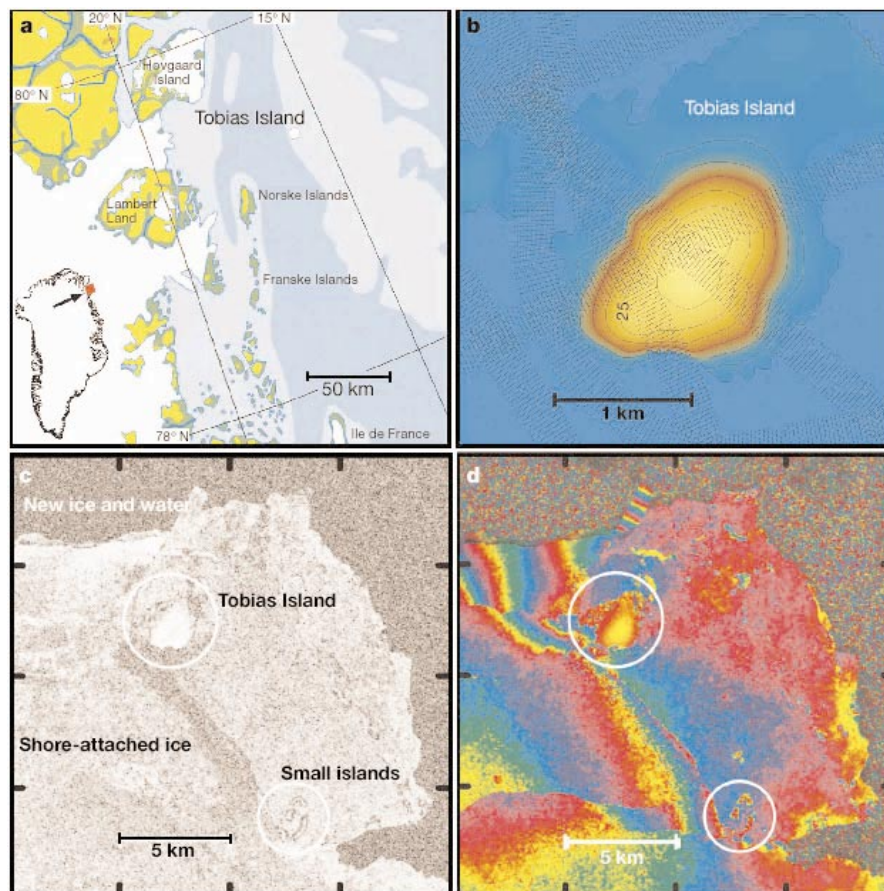


Figure 1 Tobias Island and nearby newly discovered islands that are concealed by sea ice. **a**, Map of the area (red square on inset map) showing the location of the islands off the northeastern coast of Greenland. **b**, Elevation map of Tobias Island obtained from scanning airborne laser altimetry (dotted overlay) and radar interferometry (5-m height-contour intervals). **c**, Interferometric radar correlation between two data sets obtained on 4 and 5 September 1996. Grey intensity denotes degree of correlation: black, no correlation; white, full correlation. **d**, Colour-coded interferometric phases: these are constant for land and grounded ice, but vary for floating ice. Tobias Island is located at 79° 20.3' N 15° 48.2' W; the new small islands are at 79° 15.2' N 15° 33.6' W (centre of circles, World Geodetic System 1984). Radar maps are presented in Universal Transverse Mercator projection, zone 27, with lower-left corners at 600 km east, 8,800 km north.

(100 × 100 km) east of the Nioghalvfjærd-jorden glacier clearly showed Tobias Island, but did not reveal any other bare rock islands (Fig. 1). However, inspection of interferograms from a time series of radar data acquired between December 1991 and May 1998 revealed that Tobias Island is associated with a very stable feature, which measures 0.8 × 2.0 km and is located 8 km to the southeast. We interpret this feature as either a group of islands or an area of very shallow water. It could also be a stranded iceberg, but this is less likely as its shape remained stable throughout the observation period. The radar images consistently show the presence of sea ice around Tobias Island and the 'new' islands, even during a period of break-up closer to the coast, a finding that is consistent with the idea that the semi-permanent ice cover to the south and west of Tobias Island is caused by ice adhering to various obstacles.

Satellite radar interferometry can pinpoint a location very accurately³, allowing the location to be investigated further by conventional fieldwork. For example, airborne laser altimetry was used to map Tobias Island in April 2001, in connection with the first fixed-wing aircraft landing there. Similar measurements are planned for May 2002 to verify the existence of the new islands.

Johan J. Mohr*, Rene Forsberg†

*Ørsted DTU, Technical University of Denmark, Ørsted Plads B-348, 2800 Kgs Lyngby, Denmark
e-mail: jm@oersted.dtu.dk

†National Survey and Cadastre, Rentemestervej 8, 2400 Copenhagen NV, Denmark

1. Sørensen, K. in *Geology of Greenland Survey Bulletin* (eds Dawes, P. R. & Higgins, A. K.) **186**, 7–10 (Geol. Surv. Denmark Greenland, Copenhagen, 2000).
2. Massonnet, D. & Feigl, K. *Rev. Geophys.* **36**, 441–500 (1998).
3. Mohr, J. J. & Madsen, S. N. *IEEE Trans. Geosci. Remote Sensing* **39**, 842–850 (2001).

Competing financial interests: declared none.

Palaeontology

'Modern' feathers on a non-avian dinosaur

Discoveries of integumentary coverings on non-avian theropod dinosaurs are becoming commonplace¹⁻³. But the only definitive evidence so far that any of these animals had feathers as we know them today has come from the oviraptorosaur *Caudipteryx*^{2,4,5} and the enigmatic coleurosaur *Protarchaeopteryx*², both of which are considered by some to be secondarily flightless birds^{6,7}. Here we describe the occurrence of pinnate feathers, which clearly feature a rachis and barbs, on a small, non-avian dromaeosaur from northern China. This finding indicates that feathers of modern aspect evolved in dinosaurs before the emergence of birds and flight.

The specimen, which is known as BPM 1 3-13, is a dromaeosaur skeleton housed at the Beipiao Paleontological Museum (BPM) in Liaoning province, China. It is preserved in light-grey tuffaceous sediments and was collected by the Liaoning Fossil Protection Bureau in Early Cretaceous rocks of the Jiufotang Formation⁸ at Shangheshou, near the town of Chaoyang.

Both part and counterpart of BPM 1 3-13 have been recovered (Fig. 1). The skeleton is about 95 cm long. Most of the bones are preserved in cross-section, as they shattered when the slabs split during excavation. The skull is crushed and its elements are disassociated. The head is turned backwards and the torso is twisted so that the left arm is folded back across the body and lies under the right leg. The right arm is extended; the left leg projects from the pelvis and the right leg is flexed at the knee. The long tail extends dorsally from the body (Fig. 1a).

BPM 1 3-13 can be unequivocally referred to the Dromaeosauridae on the basis of derived characters, including elongate prezygapophyses and chevrons that span several vertebrae in the tail, a retroverted pubis, and a modified second pedal digit⁹. Although BPM 1 3-13 shows several advanced maniraptoran features, such as an hourglass-shaped sternum composed of two sternal plates with attachment sulci for sternal ribs¹⁰, and a semilunate carpal that caps metacarpals 1 and 2 (ref. 10), it and other dromaeosaurs lack a reversed hallux, unserrated teeth and the short tail of basal avians. The tail is relatively much longer than that of NGMC 91 (which can probably be referred to *Sinornithosaurus*)¹¹ and that of *Microraptor*¹².

Integumentary fibres are distributed over the entire body of BPM 1 3-13. Most are not well preserved, but they show the general pattern of distribution seen in another specimen found near Lingyuan¹¹.

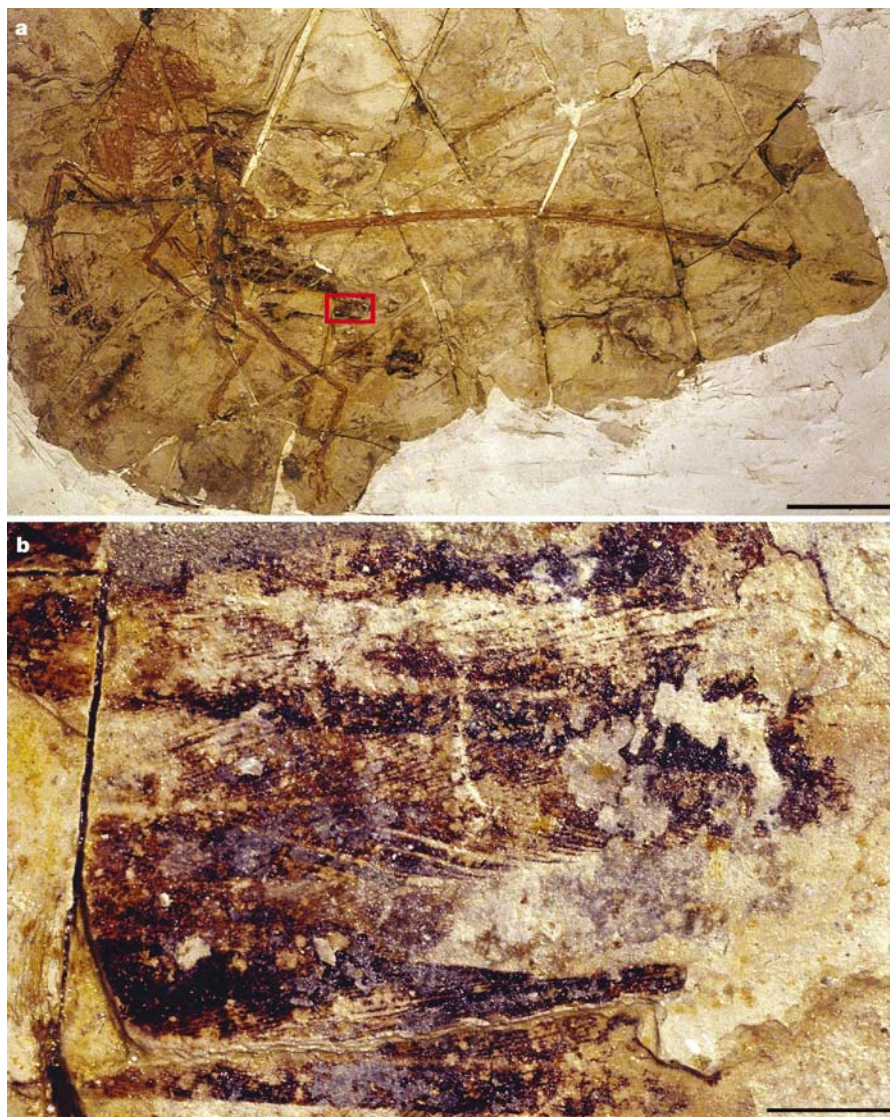


Figure 1 Evidence of modern-type feathers on the small dromaeosaur skeleton BPM 1 3-13. **a**, Slab 1, showing the whole specimen. **b**, Magnified area inside the red box in **a** showing individual feathers, each with a rachis and barbs. Scale bars: **a**, 10 cm; **b**, 50 mm.

Many of the fibres show organizational patterns similar to those in *Sinornithosaurus* which have been homologized with contour feathers³.

BPM 1 3-13 exhibits a group of feathers at the distal end of the tail (Fig. 1a). Similar groups of feathers have also been described in *Caudipteryx*². Although no fine structure can be ascertained, they extend at least 19 cm from the ultimate caudal vertebra. These feathers give BPM 1 3-13 a much more extensive tail plume than those of *Beipiaosaurus*, *Caudipteryx* and *Archaeopteryx*, although they were longer because they emanate from an area that is more proximal to the end of the tail.

Feathers of modern aspect (that is, with a rachis and barbs) are distributed along the back of the forelimbs and along the back of the hindlimbs, where they are best preserved (Fig. 1b). Feathers are also present on the front and back of the hand. The arm and leg feathers have a central rachis and symmetrical barbs similar to the remiges in

*Caudipteryx*². The hindlimb feathers extend down to the lower tibia and are more than 13.5 cm long.

The feathers of BPM 1 3-13 are structurally identical to those of modern birds, indicating not only that modern feathers must have evolved in dinosaurs before the emergence of birds and flight, but also that the feather-like structures present in many other non-avian theropods are homologous with feathers. These features may provide clues as to how feathers could have evolved from single strands to complex branching structures^{3,13}.

Mark Norell*, **Qiang Ji†**, **Keqin Gao***, **Chongxi Yuan†**, **Yibin Zhao‡**, **Lixia Wang‡**

*American Museum of Natural History, Central Park West at 79th Street, New York, New York 10024-5192, USA

e-mail: norell@amnh.org

†Institute of Geology, Chinese Academy of Geological Sciences, 26 Baiwanzhuang, Beijing 100037, China

‡Liaoning Provincial Bureau of Land and

Resources, 29 Beiling Street,
Shenyang 110032, China

- Chen, P.-J., Dong, Z.-M. & Zhen, S.-N. *Nature* **391**, 147–152 (1998).
- Ji, Q., Currie, P. J., Ji, S. & Norell, M. A. *Nature* **393**, 753–761 (1998).
- Xu, X., Zhou, Z.-H. & Prum, R. O. *Nature* **410**, 200–204 (2001).
- Sereno, P. *Science* **284**, 2137–2147 (1999).
- Norell, M. A., Clark, J. M. & Makovicky, P. J. in *New Perspectives on the Origin and Early Evolution of Birds* (eds Gauthier, J. A. & Gall, L. G.) 49–67 (Proc. Int. Symp. Hon. John H. Ostrom, Yale Peabody Mus., New Haven, Connecticut, 2001).
- Jones, T. D., Farlow, J. O., Ruben, J. A., Henderson, D. M. & Hillenius, W. J. *Nature* **406**, 716–718 (2000).
- Feduccia, A. *The Origin and Evolution of Birds* 2nd edn (Yale Univ. Press, New Haven, Connecticut, 1999).
- Jin, F. *Vert. Palas* **34**, 102–122 (1996).
- Ostrom, J. H. in *The Dinosauria* (eds Weishampel, D. B., Dodson, P. & Osmólska, H.) 269–279 (Univ. California Press, Berkeley, 1990).
- Norell, M. A. & Makovicky, P. J. *Am. Mus. Novitates* **3215**, 1–28 (1997).
- Ji, Q., Norell, M. A., Gao, K.-Q., Ji, S.-A. & Ren, D. *Nature* **410**, 1084–1088 (2001).
- Xu, X., Zhou, Z.-H. & Wang, X.-L. *Nature* **408**, 705–708 (2000).
- Prum, R. O. *J. Exp. Zool.* **285**, 291–306 (1999).

Competing financial interests: declared none.

Fibre science

Supercontraction stress in wet spider dragline

Unrestrained spider dragline ‘supercontracts’ when it is wetted, causing its length to shrink by about half and its diameter to almost double^{1,2}. Here we measure the supercontraction stresses generated upon initial exposure of spider dragline to moisture and find that they are transient, as well as being greater than previously estimated. Our findings cast doubt on suggestions that supercontraction may help to maintain tension in wet webs and could limit the potential load-bearing applications of silk and its analogues.

It has been proposed that wetting-induced supercontraction of spider dragline silk^{1–5} takes up slack in webs, restores web shape and tension after prey capture, and opposes extensional forces exerted by the weight of precipitation on webs^{1,2,6,7}, but these suggestions were based on extrapolated estimates of supercontraction stresses. However, advances in mechanical-testing instrumentation now allow supercontraction stress and its time dependence to be measured directly.

Figure 1 shows the evolution and decay of stress in a fixed length (30 mm) of *Nephila clavipes* dragline subjected to increasing humidity. Weight gain due to water absorbed by the sample makes a negligible contribution to the measured stress. The first peak in the plot of stress against time lies close to the yield strength of *N. clavipes* dragline (measured in a separate experiment at 30 °C, 60% relative humidity and a strain rate of 10^{-4} s^{-1}). A material stressed beyond its yield strength is able to undergo bulk (visco)plastic flow. If the intrinsic resistance to flow decreases with time as a result of the plasticizing effect of rising moisture content, yield will be followed by a drop in the stress/time plot (which is what we find; Fig. 1).

The ability of the stress subsequently to increase once more depends on the interplay between three factors: the relationship between moisture content and polymer-chain mobility; the rate at which moisture-enabled conformational changes can occur; and the rate at which microstructural rearrangement allows the supercontraction stress to relax. Spiders may be able to control these effects through the relative amounts of crystallizable and non-crystallizable protein in their silk⁸, thereby optimizing the fibre for use in a particular environment. Our *N. clavipes* spiders were collected in Gainesville, Florida, where the

daily humidity range at any time of year can be as large as the range in our experiments (US National Climatic Data Center, www.ncdc.noaa.gov).

The maximum supercontraction stress shown in Fig. 1 amounts to 22% of the breaking strength of the material. This result is inconsistent with estimates of the maximum supercontraction stress for dragline^{1,4}, which is quoted as no more than 4.7% of the breaking strength. Although those measurements were made using dragline from a different spider species (*Araneus diadematus*), the mechanical properties of *N. clavipes* and *A. diadematus* dragline are comparable, differing by no more than a factor of 1.5 (ref. 9). The higher maximum supercontraction stress recorded here reflects the fact that we measured instantaneous force directly and continuously as humidity increased — rather than extrapolating the force indirectly in a procedure that allows time for the supercontraction stress to relax partially.

Figure 1 also indicates that tension cannot be maintained indefinitely in a fixed length of wet dragline. After a total elapsed time of less than 5 min, the stress relaxed to zero, even though the humidity continued to rise. Although spiders might exploit supercontraction to restore the shape of a web after deformation by precipitation, wind or prey, they cannot rely on it to compensate for continued loads applied to a wet web.

Likewise, the technological applications of fibres designed to mimic spider dragline¹⁰ will be limited by stress relaxation and creep. It will be necessary to keep the fibres dry, for example by incorporating them into a water-resistant matrix¹¹, or to eliminate moisture-sensitive sequences from the protein¹².

Fraser I. Bell, Iain J. McEwen,
Christopher Viney

Department of Chemistry, Heriot-Watt University,
Edinburgh EH14 4AS, UK
e-mail: c.viney@hw.ac.uk

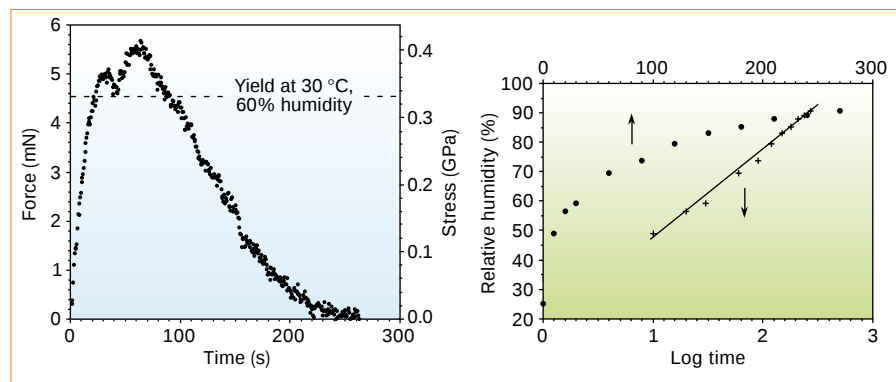


Figure 1 Time-dependent supercontraction stress in spider dragline under conditions of increasing humidity. Dragline was reeled¹³ from *Nephila clavipes* at 1 cm s^{-1} , which corresponds roughly to the rate at which silk is spun during web construction. Spiders were not anaesthetized. The average cross-sectional area of dragline (used to convert force to stress) was estimated from diameter measurements made using light microscopy¹. Forces generated by supercontraction were measured using a 10-N load cell fitted to a MicroMat 200 tensile tester (Micro Materials, Wrexham, UK). The tensile tester was housed in a benchtop environmental chamber supplied with recirculating air by an atomizing humidifier. Humidity was increased as rapidly and as greatly as the equipment would allow at the set temperature of 30 °C; the humidity increase is logarithmic over the timescale of the experiment. Further details are available from the authors.

- Work, R. W. *Textile Res. J.* **47**, 650–662 (1977).
- Work, R. W. *J. Arachnol.* **9**, 299–308 (1981).
- Gosline, J. M., Denny, M. W. & DeMont, M. E. *Nature* **309**, 551–552 (1984).
- Work, R. W. *J. Exp. Biol.* **118**, 379–404 (1985).
- Gosline, J., Nichols, C., Guerette, P., Cheng, A. & Katz, S. in *Biomimetics: Design and Processing of Materials* (eds Sarikaya, M. & Aksay, I. A.) 237–261 (Am. Inst. Phys., Woodbury, New York, 1995).
- Tillinghast, E. K. & Townley, M. in *Ecophysiology of Spiders* (ed. Nentwig, W.) 203–210 (Springer, Berlin, 1987).
- Lewis, R. V. *Acc. Chem. Res.* **25**, 392–398 (1992).
- Craig, C. L. *et al. Mol. Biol. Evol.* **17**, 1904–1913 (2000).
- Stauffer, S. L., Coughill, S. L. & Lewis, R. V. *J. Arachnol.* **22**, 5–11 (1994).
- O'Brien, J. P., Fahnestock, S. R., Termonia, Y. & Gardner, K. H. *Adv. Mater.* **10**, 1185–1195 (1998).
- Craven, J. P., Cripps, R. & Viney, C. *Composites A: Appl. Sci. Manuf.* **31**, 653–660 (2000).
- Yang, Z. *et al. J. Am. Chem. Soc.* **122**, 9019–9025 (2000).
- Thiel, B., Kunkel, D., Guess, K. & Viney, C. in *Biomolecular Materials by Design* (eds Alper, M., Bayley, H., Kaplan, D. & Navia, M.) 21–30 (Mat. Res. Soc., Pittsburgh, 1994).

Competing financial interests: declared none.

Anhydrobiosis

Plant desiccation gene found in a nematode

When subjected to drought conditions, some organisms enter a state of suspended animation known as anhydrobiosis¹, surviving for indefinite periods until rehydration allows them to resume normal metabolism. We have identified a gene in the anhydrobiotic nematode *Aphelenchus avenae* that is upregulated in response to desiccation stress and whose encoded protein shares sequence similarity with a late-embryonic gene that is induced in many plants when they are deprived of water. This finding suggests that animals and plants that undergo anhydrobiosis may use common protective strategies against dehydration, and provides a unifying insight into the mechanism of anhydrobiosis.

A characteristic feature of anhydrobiotic organisms is their synthesis of high concentrations of non-reducing sugars during the induction of anhydrobiosis¹. For example, *A. avenae* accumulates large amounts of trehalose in response to dehydration, and this correlates with its desiccation tolerance². Trehalose protects membranes and proteins from desiccation damage by replacing structural water¹, and contributes to the formation of an intracellular organic glass³ which is thought to stabilize the cell's contents. In anhydrobiotic plants, sucrose has a similar function⁴.

However, several lines of evidence indicate that non-reducing sugars alone are not sufficient to confer a state of anhydrobiosis, and that further adaptations are required^{5,6}. A class of proteins known as LEA (for 'late embryogenesis abundant') proteins accumulate in response to water deficit in many plants, and these are particularly abundant in anhydrobiotic plants such as the resurrection plant *Craterostigma plantaginum*, and in maturing seeds and pollen⁷. We therefore investigated whether similar genes are present in *A. avenae*.

We found that several genes are upregulated in *A. avenae* nematodes undergoing anhydrobiosis during exposure to 90% relative humidity for 24 hours. Of particular interest was a strongly induced transcript with a length of about 675 bases (Fig. 1a, b), which encodes a predicted protein of 143 residues. The sequence of this protein indicates that it is a member of the group-3 subclass of LEA proteins⁸; it contains at least four copies of a characteristic 11-mer motif, and further echoes of this motif are present throughout the sequence (Fig. 1c).

The group-3 LEA motif has been proposed to form an amphipathic α -helix that directs the oligomerization of the protein⁹. These proteins are extremely hydrophilic and are resistant to denaturation by heat —

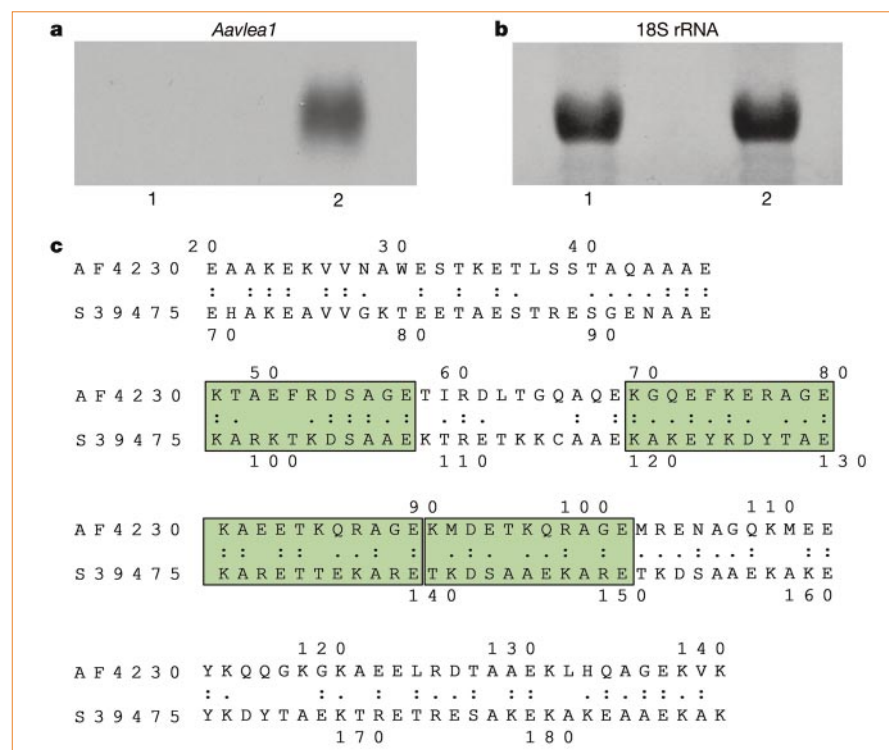


Figure 1 Desiccation induces expression of a protein in the nematode *Aphelenchus avenae* that is homologous with another embryonic protein produced by a plant under similar circumstances. **a**, Expression of the *A. avenae* late-embryonic gene *Aavlea1* in response to desiccation stress, as revealed by northern blotting using an *Aavlea1* probe: lane 1, control RNA from fresh untreated nematodes; lane 2, RNA from nematodes exposed to 90% relative humidity for 24 h. **b**, Same blot, but this time hybridized with an 18S rRNA probe as a loading control. **c**, Local alignment ($P=2e^{-19}$) of a 121-amino-acid overlap between the predicted sequences of *Aavlea1* (GenBank accession number AF423069) and an embryonic protein from the European white birch (*Betula pendula*, accession number S39475). S39475 was the most closely related LEA-3 sequence to *Aavlea1* that we were able to detect in the GenBank database. There are at least four copies of the characteristic 11-mer LEA-3 motif (highlighted).

prompting suggestions that they help to prevent damage by water stress, for example by acting as hydration buffer, molecular chaperone, ion sink or membrane stabilizer^{1,9}.

Sucrose glasses are stabilized *in vitro* by interaction with a purified group-3 LEA protein from *Typha latifolia* pollen¹⁰. Non-reducing sugars and LEA proteins may therefore act synergistically^{4,10} to promote the formation of a stable 'bioglass' in the cytoplasm of anhydrobiotic plants and in desiccation-tolerant seeds and pollen — the bioglass may trap fragile biological molecules in time and space, and preserve them from desiccation damage. If the newly identified LEA protein in *A. avenae* also stabilizes trehalose glasses in this way during desiccation, the bioglass model for anhydrobiosis can be extended to include anhydrobiotic animals.

Genomic¹¹ and partial protein¹² sequence information indicates that LEA proteins may be used by other nematodes as a defence against water stress. The genomes of some microorganisms also contain sequences that encode LEA-like proteins (although expression of these has not yet been demonstrated¹¹), including the desiccation-tolerant *Deinococcus radiodurans*, which also encodes enzymes for the biosynthesis of trehalose. A bioglass strategy for anhydrobiosis could be

widely used in biological systems and may have originated in ancient cell types that were exposed to water stress.

John Browne*, Alan Tunnacliffe†, Ann Burnell*

**Department of Biology and Institute of Bioengineering and Agroecology, National University of Ireland Maynooth, County Kildare, Ireland*

e-mail: ann.burnell@may.ie

†Institute of Biotechnology, University of Cambridge, Tennis Court Road, Cambridge CB2 1QT, UK

1. Crowe, J. H., Hoekstra, F. A. & Crowe, L. M. *Annu. Rev. Physiol.* **54**, 579–599 (1992).
2. Madin, K. A. C. & Crowe, J. H. *J. Exp. Zool.* **193**, 335–346 (1975).
3. Crowe, J. H., Carpenter, J. F. & Crowe, L. M. *Annu. Rev. Physiol.* **60**, 73–103 (1998).
4. Sun, W. Q. & Leopold, A. C. *Comp. Biochem. Physiol.* **117A**, 327–333 (1997).
5. Higa, L. M. & Womersley, C. Z. *J. Exp. Zool.* **267**, 120–129 (1993).
6. García de Castro, A. & Tunnacliffe, A. *FEBS Lett.* **487**, 199–202 (2000).
7. Ingram, J. & Bartels, D. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **47**, 377–403 (1996).
8. Dure, L. et al. *Plant Mol. Biol.* **12**, 475–486 (1989).
9. Dure, L. *Plant J.* **3**, 363–369 (1993).
10. Wolkers, W. F. et al. *Biochem. Biophys. Acta* **1544**, 196–206 (2001).
11. Dure, L. *Protein Pept. Lett.* **8**, 115–122 (2001).
12. Solomon, A., Salomon, R., Paperna, I. & Glazer, I. *Parasitology* **121**, 409–416 (2000).

Competing financial interests: declared none.

Determining the composition of the Earth

Michael J. Drake & Kevin Righter

Lunar and Planetary Laboratory, University of Arizona, Tucson, Arizona 85721-0092, USA

A long-standing question in the planetary sciences asks what the Earth is made of. For historical reasons, volatile-depleted primitive materials similar to current chondritic meteorites were long considered to provide the 'building blocks' of the terrestrial planets. But material from the Earth, Mars, comets and various meteorites have Mg/Si and Al/Si ratios, oxygen-isotope ratios, osmium-isotope ratios and D/H, Ar/H₂O and Kr/Xe ratios such that no primitive material similar to the Earth's mantle is currently represented in our meteorite collections. The 'building blocks' of the Earth must instead be composed of unsampled 'Earth chondrite' or 'Earth achondrite'.

Most workers assume implicitly that the Earth is made of some sort of extant primitive material delivered to the Earth as meteorites and probably originating in the asteroid belt. Indeed, much has been learned from meteorites about the materials present and processes occurring in the accretion disk as the planets grew. However, confusion is caused by the convenient reference of terrestrial rock abundances to CI chondrites or just 'chondrites', leading to an unintended perception that the Earth must be made of such materials.

The formation of the Earth has been debated in terms of heterogeneous accretion versus homogeneous accretion, with the former holding sway for the last 25 years or so. Heterogeneous accretion, as most prominently championed by Wänke¹, envisioned the material accreting to the Earth changing in composition and oxidation state with time. Driven by the 'stair-step' distribution of siderophile (metal-seeking) elements in the terrestrial upper mantle (Fig. 1), Wänke¹ suggested that the first 80% to 90% of the Earth accreted from very reducing materials. All elements in Fig. 1 except the refractory lithophile elements such as Sc and the rare earth elements (REE) would be quantitatively extracted into the core, and the mantle would be devoid of Fe²⁺. The next 20% to 10% or so of material accreting to the Earth would be more oxidizing, and all but the highly siderophile elements (Ir, Os, Au and so on) would remain stranded in the mantle. The highly siderophile elements were again quantitatively extracted into the core. The last roughly 1% added (the 'late veneer'²) was so oxidizing that metallic Fe did not exist (note that CI chondrites³ and Tagish Lake⁴ are the only chondrites containing no metal⁵), and all siderophile elements delivered by the 'late veneer' were forced to remain in the mantle, where they were very efficiently homogenized at the hand-specimen (centimetre) scale on a global basis. The stair-step pattern of siderophile elements in Fig. 1 is thus explained. We note that the term 'late veneer' is unfortunate, as the last dregs of material accreted to Earth are well mixed into at least the upper mantle, rather than veneering the surface. The term is, however, entrenched in the literature.

The heterogeneous accretion hypothesis makes dynamical sense in that the 'feeding zone' of the Earth must have extended further out from the Sun as the growth of planets, particularly Jupiter, increased the relative velocities and hence the eccentricities of the accreting material. However, there is a general consensus⁵ that the Earth developed one or more magma oceans late in its accretion, effectively homogenizing any pre-existing material. Thus, there is unlikely to be any record of heterogeneous accretion remaining, with the possible exception of the 'late veneer'. The bulk geochemical properties of the Earth were probably established by magmatic processes in a high-pressure and high-temperature magma ocean environment (see large blue circles in Fig. 1).

The question of whether samples of the 'building blocks' of Earth

are still extant is addressed by examining specific elemental and isotopic properties of the Earth, Mars, comets and meteorites.

Major elements in the Earth and meteorites

It has long been known that the composition of the Earth's primitive upper mantle (PUM, the Earth's mantle immediately after core formation) is distinct from that of any kind of extant meteorite. Geochemical processes on differentiated planets tend to raise the Mg/Si ratio and lower the Al/Si ratio in mantle materials from which magma has been extracted, reflecting the compatible nature of Mg and the incompatible nature of Al. Thus, Mg/Si and Al/Si ratios in samples from both Earth and Mars correlate with a negative slope (Fig. 2). In contrast, primitive materials show a loose positive correlation of unknown meaning.

Several suggestions have been made to explain the elevated Mg/Si and Al/Si ratios in the Earth's PUM relative to undifferentiated meteorites. The most prominent of these include sequestering Si in

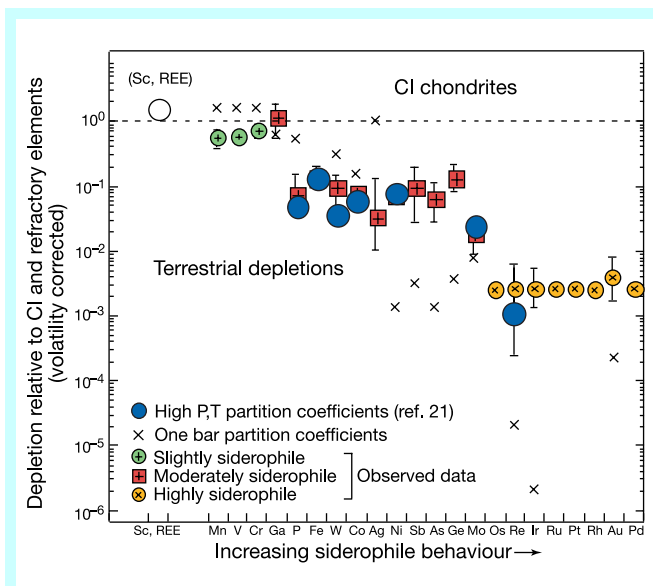


Figure 1 The abundances of elements in the Earth's primitive upper mantle after core formation show a stair-step pattern. This fingerprint contains clues to accretion and planetary differentiation. Comparison of the observed siderophile element depletions in Earth's upper mantle (green, red and yellow symbols; abundances normalized to CI carbonaceous chondrites and refractory elements), with those calculated using one bar partition coefficients (crosses) and high-pressure (P)/high-temperature (T) calculated partition coefficients (blue circles)²¹. Calculated depletions using the latter partition coefficients overlap the observed depletions, consistent with metal-silicate equilibrium and homogeneous accretion. REE, rare earth elements.

the core of the Earth, raising the Mg/Si and Al/Si ratios in the silicate mantle^{1,6}, or appealing to the possibility that the lower mantle of the Earth has a different composition from the upper mantle, that is, that PUM is not representative of the bulk mantle⁷.

Although Si can be extracted into metallic cores under very reducing conditions, it is not extracted into metal at or near the iron–wüstite oxygen buffer and at the pressures and temperatures of the base of a magma ocean in sufficient abundance to account for the high Mg/Si ratio of PUM. Further, there is no compelling experimental evidence that Si is extracted into the core under present core–mantle boundary conditions. For example, at the base of a high pressure/temperature terrestrial magma ocean, the metal/silicate partition coefficient for Si (ref. 8) is approximately 10^{-3} to 10^{-2} . It is also inconsistent with the PUM abundances of first-transition-series elements such as V, Cr and Mn, because they would be depleted far below observed abundances if conditions were sufficiently reducing to allow Si to dissolve in the Earth's core⁹.

The question of whether the composition of the lower mantle is different from the upper mantle is more complicated. If the lower mantle, constituting 68% by volume of the mantle, has a different bulk composition from PUM, then arguments about the bulk composition of the silicate Earth based on PUM are flawed. Certainly distinct radiogenic isotopic reservoirs have evolved and been preserved over time, but these systems by their very utility are intensely sensitive to geochemical fractionation events associated with mantle melting (Rb/Sr, Sm/Nd), atmospheric outgassing (I/Xe) and core segregation (U/Pb), for example. The existence of these reservoirs does not imply the existence of distinct reservoirs with radically different major-element compositions.

Indeed, most geophysical measurements are at least consistent with the upper and lower mantle having broadly the same composition. Electrical conductivity measurements on magnesiowüstite¹⁰ indicate that the lower mantle must have the same molar Mg/(Mg + Fe) ratio as the upper mantle. Ito and Takahashi¹¹

showed that the isochemical transformation of $(\text{Mg,Fe})_2\text{SiO}_4$ (spinel) with the upper-mantle molar Mg/(Mg + Fe) ratio of 0.9 to $(\text{Mg,Fe})\text{SiO}_3$ (Mg-perovskite) and $(\text{Mg,Fe})\text{O}$ (magnesiowüstite) is sharp to less than 6 km and perhaps as sharp as 1.5 km at 660 km depth, in agreement with seismic evidence for a sharp boundary. Thermal expansivity measurements¹² show that a change in chemistry (higher Fe-content) at 660 km could be compensated by an increase in temperature across the 660-km boundary, precluding evaluation of model compositions of the lower mantle using density comparisons. However, melting curves for Fe, FeO and FeS are consistent with a core–mantle boundary temperature below 3,500 K, implying that there is not a thermal boundary layer at 660 km and that the upper and lower mantle compositions are the same. Seismic tomography shows mixing of material from the upper mantle into the lower mantle and a return flow of lower mantle material into the upper mantle, as does fluid dynamical modelling¹³. Indeed, if the lower mantle was Fe-rich and had a higher density than the upper mantle, the lower and upper mantle would have to convect separately. Thus there is no compelling evidence that the bulk composition of the lower mantle is different from the upper mantle, with the possible exception of the very deep mantle¹⁴.

It is important to note that there is no such thing as a unique 'chondritic' or 'average Solar System' composition (Fig. 2). Primitive, undifferentiated meteorites, all of which presumably sample material currently present in the Main Belt of asteroids and are residues of the formation of the inner Solar System, have distinct Mg/Si versus Al/Si ratios in spite of their residence between about 2 and 4 AU. Clearly spatially associated reservoirs in the accretion disk were produced with modestly distinct bulk compositions. Estimates of the Earth's PUM are also distinct from any known type of meteorite. A reasonable conclusion is that the Earth accreted primarily from material with major-element compositions that were distinct from extant primitive meteorite types.

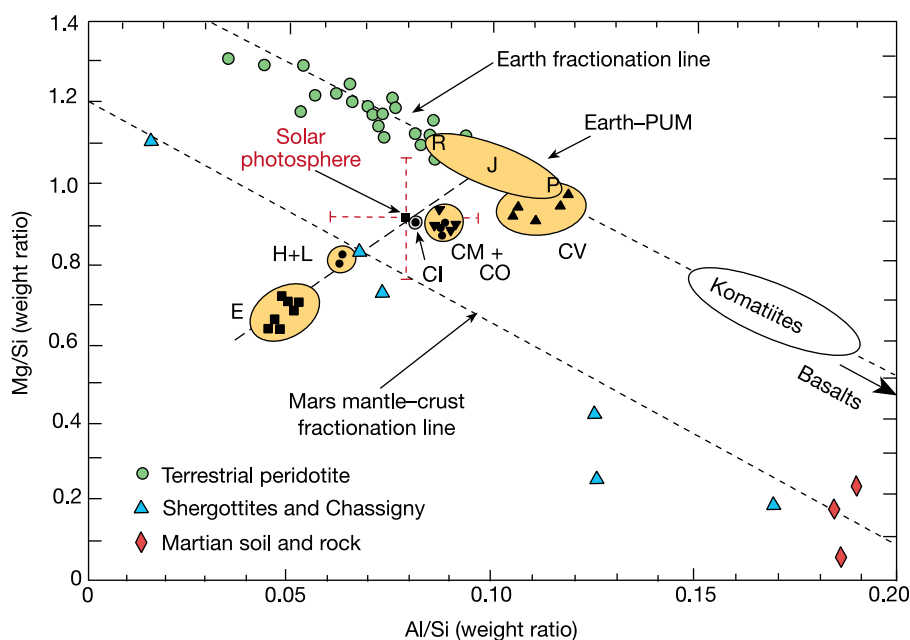


Figure 2 The major-element composition of primitive material in the inner Solar System is not of uniform composition, but defines an unexplained trend. Mg/Si versus Al/Si ratios in chondritic, terrestrial and martian materials^{35–37}. Abbreviations are as follows: enstatite (E), ordinary (H, L) and carbonaceous (CI, CM, CO and CV) chondrites. Dots are terrestrial peridotites, komatiites and basalts that fall off the diagram to the

right. The letters R, J, and P refer to estimates of the bulk silicate Earth composition (refs 38, 35 and 39, respectively). The martian fractionation line is defined by Chassigny, shergottites and martian soils and rocks from the Viking and Pathfinder mission³⁷. PUM, primitive upper mantle of the Earth.

Oxygen isotopes in the Earth

A cursory examination of the oxygen isotopic composition of meteorites indicates that only the enstatite meteorites (chondrites and achondrites) share a common oxygen reservoir with Earth, ruling out all other individual meteorite types as the 'building blocks' of the Earth (Fig. 3). Although the Earth could, in principle, be a mixture of existing meteorite types falling on both sides of the Earth–Moon–enstatite meteorite line, this possibility is unlikely. Clayton and Mayeda¹⁵ point out that the bulk oxygen compositions of the Earth, the Moon and Mars are very similar to ordinary chondrites (Fig. 3). They are sufficiently distinct from carbonaceous chondrites that no more than a few per cent of carbonaceous-chondrite-like material contributed to the formation of the Earth and terrestrial planets. Further, the existence of Mars at 1.5 AU with a different oxygen-isotopic composition from the Earth indicates that distinct oxygen reservoirs are preserved over relatively small annuli of heliocentric distance.

The existence of enstatite chondrites points to at least one primitive material with appropriate oxygen-isotopic composition. There is no reason why materials with different characteristics should not have the same oxygen-isotopic composition. Indeed, if the Earth's Moon was formed in a giant impact and the Moon is derived largely from the impactor^{16,17}, then at least two large bodies, the proto-Earth and the impactor, had identical oxygen-isotopic composition¹⁸, and are within error of the oxygen-isotopic ratio in enstatite chondrites and achondrites.

Initial $^{187}\text{Os}/^{188}\text{Os}$ ratio in PUM of the Earth

The chondritic relative abundances of highly siderophile elements (HSE) in Earth's PUM (Fig. 1) are proposed to be the result of late addition of chondritic material to the mantle after core formation had ceased (the 'late veneer'). Rhenium and osmium are two HSEs that are linked by beta decay, $^{187}\text{Re} = ^{187}\text{Os} + \beta$. Thus Os isotopes can be used as a constraint on Earth's bulk composition. The Earth's PUM has a significantly higher $^{187}\text{Os}/^{188}\text{Os}$ ratio than carbonaceous chondrites, effectively ruling them out as the source of the 'late veneer' (Fig. 4). The PUM $^{187}\text{Os}/^{188}\text{Os}$ ratio overlaps anhydrous ordinary chondrites and is distinctly higher than anhydrous enstatite chondrites. The identification of anhydrous meteorites with the

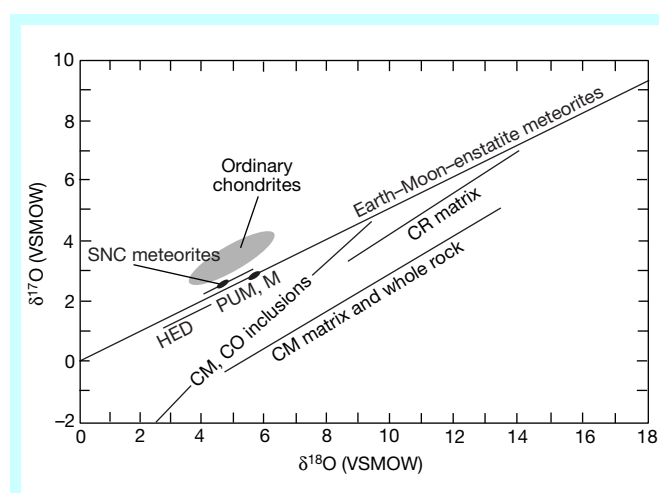


Figure 3 Simplified oxygen-isotope plot of inner Solar System material, after refs 15 and 40–42. The PUM and M (the bulk Moon) are in the small black ellipse on the Earth–Moon–enstatite meteorites line. SNC (shergottite, nakhlite, chassignite) meteorites are believed to be martian (bulk Mars denoted by black ellipse). The Solar System is divided into discrete oxygen reservoirs, and relationships between Solar System objects can be established by use of this technique. CM, CO and CR are carbonaceous chondrites. HED (howardite, eucrite, diogenite) meteorites are believed to be samples of asteroid Vesta. VSMOW, Vienna standard mean ocean water.

'late veneer' is important in the context of the possible identification of material that delivered water to Earth with the material that delivered the siderophile element 'late veneer' (see below).

D/H ratios in the Earth, Mars, meteorites and comets

The origin of Earth's water can also be investigated by examining deuterium/hydrogen (D/H) ratios in various planetary materials (Fig. 5). The D/H ratio of Vienna standard mean ocean water (VSMOW) on Earth is 150×10^{-6} , similar to the average for carbonaceous chondrites (150×10^{-6}). The corresponding average for Mars (there is considerable spread and there is no extant water ocean) is about 300×10^{-6} , similar to the value for comets: 300×10^{-6} (Fig. 5). These observations suggest local and distinct sources for most of the water in Earth and Mars.

It is possible that the D/H ratio on Mars has been raised from its original value by differential loss of H, as is suggested by the elevated D/H ratio in the martian atmosphere (Fig. 5). However, Mars has been outgassed less than the Earth because its smaller mass would lead to more rapid loss of heat. Thus, the lower martian mineral and rock D/H ratios are most probably representative of the solid undegassed planet (Fig. 5). An alternative interpretation¹⁹ notes that the lack of exchange between the martian lithosphere and hydrosphere allows for the possibility that the D/H ratio of the unsampled martian interior could be very different from rocks, minerals and atmosphere. Sampled rocks, minerals and atmosphere could reflect addition of unrecycled cometary material.

Indigenous or exogenous sources of the Earth's water

The origin of the Earth's water is intimately associated with the nature of the 'building blocks' of Earth. Two schools of thought bracket the possibilities. One view holds that temperatures were too high at 1 AU for hydrous phases to exist in the accretion disk, so that the Earth accreted 'dry'. Water, and probably organics, were most probably delivered from exogenous sources (such as comets or meteorites) after the Earth had formed. An alternative view holds that the Earth accreted 'wet', with anhydrous and hydrous silicate phases among the material accreted to the growing planet. In this view, Earth's water has an indigenous origin. We note that one version of the magma-ocean hypothesis^{20–23} requires the Earth's magma ocean to be hydrous (~3 to 4 wt% H_2O).

Delivery of water while metal is present would lead to H being extracted to the core through metal–silicate equilibrium²⁴. The

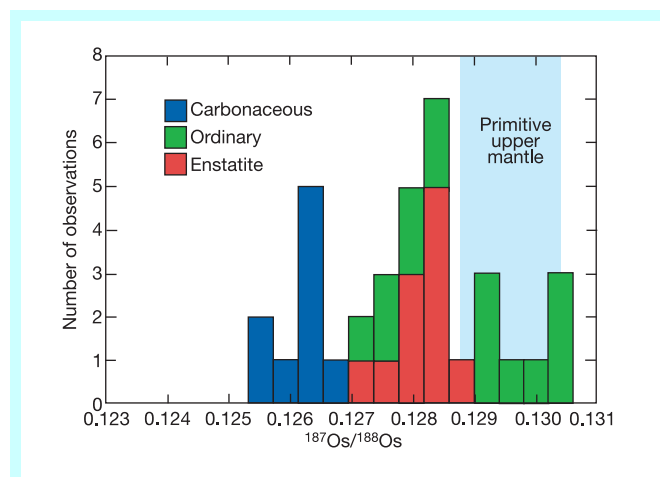


Figure 4 $^{187}\text{Os}/^{188}\text{Os}$ ratios in carbonaceous, ordinary and enstatite chondrites, and in the Earth's primitive upper mantle. The ratios are distinct and are diagnostic of the nature of the Earth's 'late veneer'⁴³. Mars is not plotted because the uncertainty in its initial $^{187}\text{Os}/^{188}\text{Os}$ ratio is larger than the range of the x axis.

efficiency of this process is not well characterized, but work until now suggests that Earth must have had large additions of water during accretion so that there was enough left in the mantle to make the oceans. Given such water additions, there are a number of possible ways in which the Earth received its water:

(1) Earth could have accreted 'dry' with all water being added late in accretion from carbonaceous chondrites, or 'wet' with some addition of water from carbonaceous chondrites (exogenous). In this case, the amount of carbonaceous chondrite material added would scale directly with the mass of water in the Earth. However, as discussed below, neither scenario is consistent with the observation that the HSE 'late veneer' and the Earth's Kr/Xe ratio are not carbonaceous chondritic.

Late delivery of water from carbonaceous chondritic material must be consistent with the siderophile element 'late veneer' (see Fig. 1). Let us make conservative assumptions designed to maximize the possibility that late-delivered asteroidal material is the source both of the Earth's water and the 'late veneer'. The mass of the Earth is 5×10^{27} g and the mass of the Earth's mantle is 4.11×10^{27} g. The minimum mass of water in the Earth is the current water content of the Earth's oceans, about 1.4×10^{24} g. We note, however, water masses up to 50 times this amount have been postulated²⁵. The abundance level of siderophile elements which characterize the 'late veneer' is $0.003 \times \text{CI}$ (Fig. 1), where CI is the abundance of siderophile elements in CI carbonaceous chondrites. It has been suggested that the abundance level might be as high as $0.007 \times \text{CI}$ (ref. 26), but this factor-of-two uncertainty is insignificant for the arguments that follow. The mass of material necessary to generate the siderophile element 'late veneer' is about 10^{25} g if the 'late veneer' is mixed throughout the entire mantle of the Earth, and about 0.3×10^{25} g if the 'late veneer' is mixed only into the upper mantle.

Using these numbers, assuming carbonaceous chondritic material containing 10 wt% water, and assuming that the fraction of water retained by Earth is 1, the mass of material accreted to yield the minimum mass of water in the Earth is 1.4×10^{25} g. This value is not distinguishable from the mass of material needed to generate the 'late veneer' in the whole mantle, 10^{25} g, but appears to be a factor of five too high if the 'late veneer' is only mixed into the upper mantle (0.3×10^{25} g)²⁶. Moreover, addition of such a water-rich 'late veneer' would require carbonaceous chondritic material, but the Os-isotopic compositions of carbonaceous chondrites are inconsistent with that of the Earth's PUM.

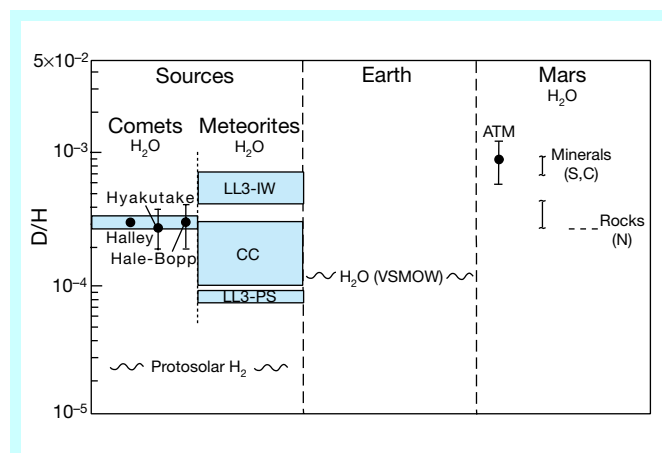


Figure 5 The D/H ratios in H₂O in three comets, meteorites, Earth, protosolar H₂, and Mars^{19,28,29,44–46}. CC, carbonaceous chondrites; LL3-IW, interstellar water in Semarkona; LL3-PS, protostellar water in Semarkona. D/H, deuterium/hydrogen ratio. ATM, atmosphere.

Recall, further, that we have made the most conservative assumptions. C-type asteroids are concentrated in the outer part of the Main Belt. CI, CM and CR carbonaceous chondrites contain 3,600–11,800 p.p.m. H (ref. 27), corresponding to about 3–9 wt% H₂O if all the H is in H₂O. It is unclear how much of the H measured is bound up in organic material rather than H₂O. Thus, the average H₂O-content of these carbonaceous chondrites may be much less than the maximum value of about 9 wt% H₂O. CO carbonaceous chondrites contain trivial amounts of H. Other asteroid types, such as S-types, are concentrated in the inner part of the belt and are anhydrous. If the average water content of the accreting material is lower than 10 wt%, as it must have been because other lower water-content or anhydrous material must also have been accreted, then the mass of material required to generate the minimum mass of water in the Earth would increase inversely with the average water content of accreting material. If the fraction of water retained is less than 1, as seems probable, the mass required also increases by inverse relation. If the 'late veneer' is restricted to the upper mantle, only, then the mass required to deliver the minimum mass of the Earth's water is three times higher than that required to deliver the 'late veneer'. If the mass of the Earth's water exceeds the minimum (one Earth ocean mass) used here, as seems inevitable, the mass of material required to deliver that water would exceed the mass required by the 'late veneer' by a precisely proportionate amount.

Stated differently, there is equivalence between the mass of the 'late veneer' and the mass of water only if the hydrous material averages 10 wt% water, the fraction of water retention is one, the mass of water in the Earth corresponds to one Earth ocean (the minimum mass), the 'late veneer' is mixed throughout the mantle, and currently unsampled hydrous material with PUM isotopic ratios was accreted.

(2) Earth could have accreted 'wet' with some exogenous addition of water from comets. Indigenous Earth water could have had D/H ratios representative of the inner Solar System, that is, low values because of relatively higher nebular temperatures, perhaps like protosolar hydrogen ($2\text{--}3 \times 10^{-5}$)²⁸, in which case a cometary contribution of up to 50% is possible. Alternatively, indigenous Earth water could have had D/H ratios representative of a protostellar water component identified in meteorites ($\sim 9 \times 10^{-5}$)²⁹, in which case there could be as little as a 10–15% cometary contribution¹⁹.

Delivery of water from comets can be evaluated in the light of other cometary geochemical data. For instance, for an assumed Ar/H₂O ratio of 1.2×10^{-7} in the bulk Earth, comets like Hale-Bopp with an approximately solar ratio of Ar/H₂O (ref. 30) would bring in 2×10^4 more Ar than is presently in the Earth's atmosphere³¹, if 50% of Earth's water, the maximum amount permitted by D/H ratios, was derived from comets. If this measurement of comet Hale-Bopp is applicable to all comets (note that an Ar/O ratio of less than 0.1 times the solar ratio has recently been reported for comet LINEAR³²) that collided with Earth over Solar System history, then there has been minimal delivery of water in the form of comets subsequent to core formation. Both Ar/H₂O and D/H data suggest either that (1) the 'late veneer' was non-cometary water-bearing material or that (2) Earth accreted 'wet'.

(3) Earth could have accreted 'wet', that is, from hydrous and anhydrous materials, with no significant addition of water from exogenous sources. In this case, hydrous phases would have a mean D/H ratio yielding VSMOW in liquid water. Oxygen, D/H and Os isotopic ratios all point to this scenario being most likely, as these ratios rule out extant meteoritic materials as sources of the Earth's water. If the Ar/H₂O ratio measured in comet Hale-Bopp is correct and representative of all comets colliding with Earth (again, see ref. 32), then cometary water may also be ruled out. The most probable source of water is a hydrous inner Solar System reservoir at approximately 1 AU, but we cannot formally rule out a single wet planetary embryo from the asteroid belt or a large comet with the

appropriate elemental and isotopic characteristics³³. Such an asteroid or comet would have to be unlike any sampled materials, however.

Kr/Xe ratios

⁸⁴Kr/¹³⁰Xe ratios in the Earth's atmosphere, average Solar System material, CI carbonaceous chondrites, and enstatite chondrites³⁴, are given in Table 1. It is immediately clear that Earth's noble gases are distinct from both types of primitive meteorites. These differences may have arisen from some secondary process. Regardless, it seems unlikely that Earth's Kr/Xe ratio is derived from carbonaceous chondrites.

What the Earth is not made of

Selected features of the Earth, Mars, primitive meteorites, and comets imply mutually inconsistent requirements for the Earth to be composed of extant materials. For example:

(1) The major-element composition of the PUM of the Earth seems to be part of a progression of closely related compositions (Fig. 2). There is a continuum of planetary materials increasing in Mg/Si and Al/Si ratios in the order: enstatite, ordinary, bulk Mars, carbonaceous chondrites, PUM.

(2) Oxygen isotopes require that the Earth be made of either enstatite meteorites or a mixture of meteorites falling above and below the Earth–Moon–enstatite meteorite mass fractionation line (Fig. 3). The similarity of the bulk oxygen-isotopic composition of the Earth and chondritic meteorites precludes more than a few per cent of carbonaceous chondritic material accreting to the Earth (Fig. 3). Further, the existence of Mars at 1.5 AU with a different oxygen-isotopic composition from the Earth points to distinct oxygen reservoirs being preserved over relatively small annuli of heliocentric distance.

(3) The Os-isotopic composition of the PUM of the Earth, believed to be that of a late addition called the 'late veneer', is inconsistent with water-bearing carbonaceous chondrites, but is consistent with anhydrous ordinary chondrites (Fig. 4).

(4) Measurements of D/H ratios in Earth and carbonaceous chondrites differ from Mars and comets by about a factor of two (Fig. 5). These observations are best explained by an indigenous and distinct source of most of the water on Earth.

(5) Measurements of ⁸⁴Kr/¹³⁰Xe ratios in Earth's atmosphere are different from extant carbonaceous and ordinary chondritic meteorites.

(6) A source of water further from the Sun than 1 AU requires water-bearing carbonaceous chondrites to deliver water. Mass balance considerations, however, indicate that even the minimum water budget of the Earth being delivered in this manner would lead to excesses of highly siderophile elements in the PUM relative to observation, if the 'late veneer' is mixed only into the upper mantle. Comets can deliver no more than 50 wt% of the Earth's minimum water budget.

What the Earth is made of

The logical conclusion of the previous discussion of major element, O and Os isotope, noble gas and D/H data is that the Earth accreted at least in part from hydrous materials that are not represented in

our meteorite collections—they are no longer extant in the inner Solar System. Free water in the Earth is derived from such material accreted after core formation effectively ceased. A cometary contribution of less than 50 wt% of the Earth's minimum water budget cannot be presently excluded.

In this case, the similarity of the ¹⁸⁷Os/¹⁸⁸Os ratio in PUM with ordinary chondrites and oxygen isotopic composition of PUM with the enstatite meteorites is *prima facie* evidence for the existence of Earth-building materials sharing some properties in common with various extant meteorites, although no extant meteorites share all of the properties of Earth material. All such material is either contained in the Earth and Moon or was ejected from the inner Solar System, consistent with the fate of material during the giant impact that is believed to have formed the Moon. Earth-forming material contained both metal and silicate, at least some portion of which was hydrous.

Although comets and known meteorite types are not suitable candidates for exogenous sources for most of Earth's water, we note that many potential exogenous, water-bearing materials have not yet been characterized or recognized. Studies of Kuiper-belt objects, trans-neptunian objects, interplanetary dust particles (IDPs), interstellar ice, and rare meteorites may all offer new insights into this problem. Only if we postulate the existence of hydrous material with ordinary chondritic Os isotopes, a material not currently sampled on Earth, can the 'late veneer' be reconciled with the Earth's water budget.

A unique bulk composition for Earth implies that compositionally distinct reservoirs were maintained in the main asteroid belt and in planetesimals at Earth's and Mars' semi-major axes. The distinct oxygen-isotope signatures of Earth and Mars reinforce that conclusion. The implication is that there was not widespread mixing of the bulk of the material in the inner Solar System during accretion and that planetary objects forming in the inner Solar System received the overwhelming majority of their masses from narrow annuli around the Sun. □

Table 1 Ratios of ⁸⁴Kr/¹³⁰Xe in various Solar System objects³⁴

Object	⁸⁴ Kr/ ¹³⁰ Xe ratio
Solar System	0.7×10^{-2}
CI carbonaceous chondrites	5.1×10^{-2}
Enstatite chondrites	(5–23)
Earth's atmosphere	12×10^{-2}

We note that extant meteorites have ⁸⁴Kr/¹³⁰Xe ratios that are very different from Earth's atmosphere.

- Wänke, H. Constitution of terrestrial planets. *Phil. Trans. R. Soc. Lond.* **303**, 287–302 (1981).
- Chou, C.-L. Fractionation of siderophile elements in the Earth's upper mantle. *Proc. Lunar Planet. Sci. Conf.* **9**, 219–230 (1978).
- Brearely, A. J. & Jones, R. H. in *Planetary Materials. Reviews in Mineralogy* Vol. 36 (ed. Papke, J. J.) 3–1–3–398 (The Mineralogical Society of America, Washington DC, 1998).
- Brown, P. G. *et al.* The fall, recovery, orbit, and composition of the Tagish Lake meteorite: a new type of carbonaceous chondrite. *Science* **290**, 320–325 (2000).
- Drake, M. J. Accretion and primary differentiation of the Earth: a personal journey. *Geochim. Cosmochim. Acta* **64**, 2363–2370 (2000).
- Allegre, C. J., Poirer, J.-P., Humler, E. & Hofmann, A. W. The chemical composition of the Earth. *Earth Planet. Sci. Lett.* **134**, 515–526 (1995).
- Anderson, D. L. Composition of the Earth. *Science* **243**, 367–370 (1989).
- Wade, J. & Wood, B. J. The Earth's 'missing' niobium may be in the core. *Nature* **409**, 75–78 (2001).
- Drake, M. J., Newsom, H. E. & Capobianco, J. C. V. Cr, and Mn in the Earth, Moon, EPB, and the origin of the Moon: Experimental studies. *Geochim. Cosmochim. Acta* **53**, 2101–2111 (1989).
- Wood, B. J. & Nell, J. High-temperature electrical conductivity of the lower-mantle phase (Mg,Fe)O. *Nature* **351**, 309–311 (1991).
- Ito, E. & Takahashi, E. Post-spinel transformations in the system Mg₂SiO₄–Fe₂SiO₄ and some geophysical implications. *J. Geophys. Res.* **94**, 10637–10646 (1989).
- Chopelas, A. & Boehler, R. Thermal expansivity in the lower mantle. *Geophys. Res. Lett.* **19**, 1983–1986 (1992).
- Tackley, P. J., Stevenson, D. J., Glatzmaier, G. A. & Schubert, G. Effects of an endothermic phase transition at 670 km depth in a spherical model of convection in the Earth's mantle. *Nature* **361**, 699–704 (1993).
- Kellogg, L. H., Hager, B. H. & Van der Hilst, R. D. Compositional stratification in the deep mantle. *Science* **283**, 1881–1884 (1999).
- Clayton, R. N. & Mayeda, T. K. Oxygen isotope signatures of hydration reactions in solar system materials. *11th Annu. V. M. Goldschmidt Conf. Abstr.* no. 3648, LPI Contribution No. 1088 (Lunar and Planetary Institute, Houston, 2001).
- Cameron, A. G. W. in *Origin of the Earth and Moon* (eds Canup, R. M. & Righter, K.) 133–144 (Univ. Arizona Press, Tucson, 2000).
- Canup, R. M. & Asphaug, E. The Moon-forming impact. *Nature* **412**, 708–712 (2001).
- Weichert, U. *et al.* Oxygen isotope homogeneity of the Moon. *Lunar Planet. Sci. Conf.* **32**, Abstr. no. 1669, LPI Contribution No. 1080 (Lunar and Planetary Institute, Houston, 2001).
- Owen, T. & Bar-nun, A. in *Origin of the Earth and Moon* (eds Canup, R. M. & Righter, K.) 459–471 (Univ. Arizona Press, Tucson, 2000).
- Righter, K., Drake, M. J. & Yaxley, G. Prediction of siderophile element metal/silicate partition

- coefficients to 20 GPa and 2800 °C: the effects of pressure, temperature, oxygen fugacity, and silicate and metallic melt compositions. *Phys. Earth Planet. Inter.* **100**, 115–134 (1997).
21. Righter, K. & Drake, M. J. Metal/silicate equilibrium in a homogeneously accreting Earth: New results for Re. *Earth Planet. Sci. Lett.* **146**, 541–554 (1997).
 22. Righter, K. & Drake, M. J. Effect of water on metal/silicate partitioning of moderately siderophile elements: a high pressure and temperature terrestrial magma ocean and core formation. *Earth Planet. Sci. Lett.* **171**, 383–399 (1999).
 23. Righter, K. & Drake, M. J. Metal/silicate equilibrium in the early Earth—new constraints from volatile moderately siderophile elements. *Geochim. Cosmochim. Acta* **64**, 3581–3507 (2000).
 24. Okuchi, T. Hydrogen partitioning into molten iron at high pressure: implications for Earth's core. *Science* **278**, 1781–1784 (1997).
 25. Abe, Y., Ohtani, E., Okuchi, T., Righter, K. & Drake, M. J. in *Origin of the Earth and Moon* (eds Canup, R. M. & Righter, K.) 413–433 (Univ. Arizona Press, Tucson, 2000).
 26. Morgan, J. W., Walker, R. J., Brandon, A. D. & Horan, M. F. Siderophile elements in earth's upper mantle and lunar breccias: data synthesis suggests manifestations of the same late influx. *Meteor. Planet. Sci.* **36**, 1257–1275 (2000).
 27. Zinner, E. in *Meteorites and the Early Solar System* (ed. Kerridge, J.) 956–983 (Univ. Arizona Press, Tucson, 1988).
 28. Lecluse, C. & Robert, F. Hydrogen isotope exchange reaction rates: origin of water in the inner solar system. *Geochim. Cosmochim. Acta* **58**, 2927–2939 (1994).
 29. Deloule, E. & Robert, F. Instellar water in meteorites? *Geochim. Cosmochim. Acta* **59**, 4695–4706 (1995).
 30. Stern, S. A. *et al.* The discovery of argon in Comet C/1995 O1 (Hale-Bopp). *Astrophys. J.* **544**, L169–L172 (2000).
 31. Swindle, T. D. & Kring, D. A. Implications of noble gas budgets for the origin of water in Earth and Mars. *11th Annu. V. M. Goldschmidt Conf.* Abstr. no. 3785, LPI Contribution No. 1088 (Lunar and Planetary Institute, Houston, 2001).
 32. Feldman, P. D., Weaver, H. A. & Burgh, E. B. FUSE observations of CO and H₂ emission in comet C/2001 A2 (LINEAR). *Bull. Am. Astron. Soc.* **33**, 1120 (2001).
 33. Morbidelli, A. *et al.* Source regions and timescales for delivery of water to the Earth. *Meteor. Planet. Sci.* **35**, 1309–1320 (2000).
 34. Pepin, R. O. in *Origin and Evolution of Planetary and Satellite Atmospheres* (eds Atreya, S. K., Pollack, J. B. & Matthews, M. S.) 291–305 (Univ. Arizona Press, Tucson, 1989).
 35. Jagoutz, E. *et al.* The abundances of major, minor and trace elements in the earth's mantle as derived from primitive ultramafic nodules. *Proc. Lunar Planet. Sci. Conf.* **10**, 2031–2050 (1979).
 36. Taylor, S. R. *Solar System Evolution: A New Perspective* (Cambridge Univ. Press, Cambridge, 1992).
 37. Dreibus, G. *et al.* Relationship between rocks and soil at the Pathfinder landing site and the martian meteorites. *Lunar Planet. Sci. Conf.* [CD-ROM] **29**, Abstr. no. 1348 (Lunar and Planetary Institute, Houston, 1998).
 38. Ringwood, A. E. *Origin of the Earth and Moon* (Springer, New York, 1979).
 39. Palme, H. & Nickel, K. G. Ca/Al ratio and composition of the Earth's primitive upper mantle. *Geochim. Cosmochim. Acta* **49**, 2123–2132 (1986).
 40. Clayton, R. N. & Mayeda, T. K. Oxygen isotope studies in carbonaceous chondrites. *Geochim. Cosmochim. Acta* **63**, 2089–2104 (1999).
 41. Clayton, R. N. Oxygen isotopes in meteorites. *Annu. Rev. Earth Planet. Sci.* **21**, 115–149 (1993).
 42. Clayton, R. N., Mayeda, T. K., Goswami, J. N. & Olsen, E. J. Oxygen isotope studies in ordinary chondrites. *Geochim. Cosmochim. Acta* **55**, 2317–2337 (1991).
 43. Meisel, T., Walker, R. J., Irving, A. J. & Lorand, J.-P. Osmium isotopic compositions of mantle xenoliths: a global perspective. *Geochim. Cosmochim. Acta* **65**, 1311–1323 (2001).
 44. Robert, F., Gautier, D. & Dubrulle, B. The solar system D/H ratio: Observations and theories. *Space Sci. Rev.* **92**, 201–224 (2000).
 45. Deloule, E., Robert, F. & Doukhan, J. C. Interstellar hydroxyl in meteoritic chondrules: implications for the origin of water in the inner solar system. *Geochim. Cosmochim. Acta* **62**, 3367–3378 (1998).
 46. Robert, F. The origin of water on Earth. *Science* **293**, 1056–1058 (2001).

Acknowledgements

Supported by NASA and NSF. We thank D. Lauretta for comments. A review by T. Owen brought noble gas ratios in Solar System bodies to our attention. Discussions with H. McSweeney, D. Kring, R. Boehler, D. Mao, L. Stixrude, A. Morbidelli, J. Lunine, F. Robert and T. Swindle have been helpful.

Correspondence and requests for materials should be addressed to M.J.D (e-mail: drake@lpl.arizona.edu).

Out of Africa again and again

Alan R. Templeton

Department of Biology, Washington University, St Louis, Missouri 63130-4899, USA

The publication of a haplotype tree of human mitochondrial DNA variation in 1987 provoked a controversy about the details of recent human evolution that continues to this day. Now many haplotype trees are available, and new analytical techniques exist for testing hypotheses about recent evolutionary history using haplotype trees. Here I present formal statistical analysis of human haplotype trees for mitochondrial DNA, Y-chromosomal DNA, two X-linked regions and six autosomal regions. A coherent picture of recent human evolution emerges with two major themes. First is the dominant role that Africa has played in shaping the modern human gene pool through at least two—not one—major expansions after the original range extension of *Homo erectus* out of Africa. Second is the ubiquity of genetic interchange between human populations, both in terms of recurrent gene flow constrained by geographical distance and of major population expansion events resulting in interbreeding, not replacement.

Recent human evolution is an area of great controversy¹. There is general agreement that the human lineage evolved in Africa, and then spread to southern Eurasia as *Homo erectus*. After *Homo erectus* spread out of Africa, the out-of-Africa replacement model² posits that populations in Africa, Europe and Asia had little genetic contact and evolved independently, with anatomically modern humans evolving only in Africa. After the evolution of modern humans in Africa, a second expansion occurred out of Africa about 100,000 years ago that resulted in the global replacement and genetic extinction of nonmodern human populations by anatomically modern humans. Under the multiregional trellis model^{3,4} genetic contact between African and non-African *Homo erectus* populations was maintained although restricted by isolation by distance. Isolation by distance allowed local populations to become differentiated from one another, but gene flow prevented long-term independent evolution such that humanity evolved into modernity as a single evolutionary lineage.

Much of this controversy started with the publication of evolutionary trees of haplotype variation in human mitochondrial DNA (mtDNA)^{5,6}. The estimation of these initial mtDNA haplotype trees was flawed^{7,8}, but many subsequent studies have generated more reliable haplotype trees for mtDNA^{9–12}, Y-chromosomal DNA regions (Y-DNA)¹³, X-linked DNA regions^{14,15} and several autosomal DNA regions^{16–20}. Not all of these haplotype trees have been analysed with respect to models of recent human evolution. The purpose of this paper is to analyse several recently published haplotype trees and combine these new analyses with older analyses.

Reconstructing human evolution without previous models

Instead of testing *a priori* hypotheses of recent human evolution, nested clade phylogeographic analyses²¹ as implemented with the program GEODIS²² are used to infer both historical events (such as range expansions) and recurrent events (such as gene flow) without regard to any prior model. Although it is relatively new²¹, nested clade analysis has been validated by applying it to data sets with known *a priori* information and has been found to be accurate and not prone to false positives, although it fails to detect all known events²³. This methodology has now become a standard tool in phylogeographic analyses (for example, see refs 24–27).

The nested clade analysis first tests the null hypotheses of no association between geography and the haplotype tree. Only when this null hypothesis is rejected at the 5% level of significance is there an attempt to interpret the pattern biologically. Inferences are thereby limited to patterns that are based upon sufficient sampling and genetic resolution to have a significant phylogeographic signal. The biological meaning of these significant signals is interpreted by using an inference key²¹. This key can lead to the conclusion that the

sampling design was inadequate to make clear biological inferences. Therefore, only a subset of the statistically significant results lead to a biological interpretation. The use of an *a priori* inference key published well before the production of all the data sets to be analysed here prevents me from making *post hoc* explanations or trying to fit the results to a favoured hypotheses. The GEODIS program and the inference key are available at http://biog.byu.edu/zoology/crandall_lab/geodis.htm. The key, the GEODIS program documentation, and a fuller description of the nested clade methodology are available as Supplementary Information.

In choosing data for analyses from the recent literature, four criteria were used: (1) a minimum of four populations had to be sampled with at least one each in Europe, Asia and Africa; (2) the minimum sample size was 35 individuals; (3) haplotypes were determined, and the frequencies of these haplotypes within each sampled population were available; and (4) there was little or no recombination within the DNA region being investigated, allowing the estimation of a biologically meaningful haplotype tree. The data sets satisfying these criteria include a recent survey of complete mtDNA genome variation in humans⁹, the X-linked gene coding for the pyruvate dehydrogenase *E1* α -subunit (*PDHA1*)¹⁴, and five autosomal regions: these are the *MX1* locus on chromosome 21 (ref. 17), the gene coding for eosinophil-derived neurotoxin (*EDN*) on chromosome 14 (ref. 20), the gene coding for eosinophil cationic protein (*ECP*) on chromosome 14 (ref. 20), the gene coding for the melanocortin 1 receptor (*MCR1*) on chromosome 16 (ref. 18), and compound haplotypes generated in the MS205 hypervariable minisatellite region on chromosome 16 (ref. 19). The whole genome mtDNA survey estimated a haplotype tree that excluded the sites in the D-loop region to focus upon the older phylogenetic signal in the mtDNA⁹, and the current analysis also excludes the D-loop. Haplotype trees were estimated by statistical parsimony²⁸ using the program TCS²⁹. In all cases, the statistical parsimony tree is either identical to or is a topologically compatible but more resolved version of the haplotype tree estimated by the original authors, who usually used either maximum parsimony or neighbour-joining. In some cases, tree ambiguities were further resolved using coalescent criteria³⁰.

The results obtained from these DNA regions are combined with previously published nested clade analyses^{13,23,31–33} of mtDNA data^{10–12,34,35}, the *SRY* and *YAP* regions on the Y chromosome¹³, a non-coding, low-recombination region on the X chromosome called Xq13.3 (ref. 15), and the autosomal haemoglobin β -chain locus¹⁶.

Results

The biological inferences made from previously published nested

Table 1 Previously published inferences using the GEODIS program

DNA region	Mitochondrial DNA			SRY and YAP	Xq13.3	Haemoglobin β
Reference Range	36*, 37*, 33† Global	11*, 12*, 23† NE Asia and America	10*, 23† SE Asia and Pacific	13* Global	15*, 35† Global	16*, 34† Global
	IBD	IBD, LDC to America	IBD, LDD	RE out of Africa		RE out of Africa
	IBD	IBD, RE to Siberia	LDC to Pacific	IBD, RE out of Africa	IBD	IBD
			IBD, RE and/or LDC to Pacific			
		F (Asia versus America), IBD				
	IBD; RE to N Eurasia	IBD, LDD, RE within America	IBD, RE to Pacific	IBD, RE out of Asia	IBD	RE out of Asia

Inferences are shown from oldest (top) to most recent (bottom). IBD, recurrent gene flow restricted by isolation by distance; LDD, recurrent gene flow with some long-distance dispersal; RE, range expansion of populations to geographically contiguous areas; LDC, range expansion of populations through long-distance colonization; F, genetic fragmentation into two or more allopatric populations.

* Reference for the data.

† Reference for the analysis (if different).

clade analyses are summarized in Table 1. The geographical extent of sampling is indicated, with 'global' meaning that human populations in Africa, Europe, Asia and America were included. Only statistically significant (5% level) inferences are shown that also resulted in an unambiguous biological interpretation. The inferences are ordered from the highest clade level in the nested analysis (the oldest contrasts in the haplotype tree) to the lowest clade level (the most recent contrasts). When no inference is indicated on the first inference line in Table 1, either no statistically significant associations were detected at the highest nesting level or their biological meaning was ambiguous. Similarly, if no entry occurs on the bottom line, no statistically significant, biologically unambiguous inference occurred at the most recent level of analysis. Inferences are ordered solely with respect to the nesting structure for a specific gene or DNA region. Table 1 does not give a temporal ordering of inferences across genes.

The output from even a single GEODIS analysis is lengthy, so the results of the new analyses are not given here but are available in the Supplementary Information. Table 2 presents the unambiguous biological inferences associated with statistically significant phylogeographic associations, with relative temporal ranks within each gene.

Integrating inferences across loci

There is no expectation that every haplotype tree should yield the same inferences. The inferences obtained from any one tree depend upon several sampling features, which are summarized in Table 3 for the data sets analysed. Table 3 reflects the numbers used in the actual GEODIS analyses, which may differ from the numbers given in the original references. For example, the genetic survey of a non-coding region on the X chromosome was performed on 69 individuals, each from a different location¹⁵. Having a sample size of one at every location precludes any assessment of within-population variation versus between-populations variation, so several nearby populations were pooled to reduce the total number of geographical sites to 29 for the purposes of the nested phylogeographic analysis³³.

One sampling constraint on inference is the spatial spread of the

samples relative to the distribution of the species. Not all data sets covered the same geographic areas, and the inferences from any one data set are confined to the area sampled. Another constraint is the number of locations sampled, which varies from 4 to 35 (Table 3). For example, the analysis of the mtDNA of Torroni *et al.*^{11,12} was unique in detecting range expansion, long-distance dispersal, and isolation by distance within the Americas. Most of the other data sets had only one native American population, thereby making inferences within the American continents impossible. As the number of locations within an area increases, the spatial scale of inference becomes finer.

A third aspect of sampling is sample size, which varies from 35 to 1,544 (Table 3). Biological inference is confined to statistically significant associations, so the number of individuals sampled is important. In general, the larger the sample size, the more power in making inferences. Also, there is generally more power to detect older events or processes that influence many populations and haplotypes than more recent events or forces. For example, the smallest sample in Table 3 (the X-linked *PDHA1* locus) could only detect a statistically significant signal at the highest level of nesting, a contrast that makes use of all 35 individuals sampled.

The DNA region sampled also determines inference properties. For example, only mtDNA detected the fragmentation event between native American populations and Old World (Asian) populations (Table 1). This event was too recent (perhaps only 14,000 years ago) to have much chance of being marked by mutations in the other genetic systems, owing to their slower rates of overall evolution. Moreover, major range expansions can occur that are not detected by a particular locus simply because mutations did not occur at the right time and place to mark the event²³. Thus, although the nested analysis is not prone to false positives, any one haplotype tree may not detect certain events or processes owing to inadequate genetic resolution²³. Genetic resolution also constrains what is meant by a 'recurrent' evolutionary process²¹. For example, when gene flow is inferred, it does not necessarily mean that gene flow occurred every generation, but rather that gene flow occurred recurrently relative to the mutation

Table 2 New inferences using the GEODIS program

DNA region	mtDNA	PDHA1	MS205	MC1R	MX1	EDN	ECP
Reference Range	9 Global	14 Africa and Eurasia	19 Global	18 Global	17 Global	20 Global	20 Global
	RE out of Africa	IBD	RE, origin ambiguous RE to Pacific and America	RE out of Africa	IBD RE to Pacific	IBD	IBD
	F (America)		IBD, RE to Pacific	RE to N Eurasia	IBD, RE to America	RE to America	

Inference abbreviations are the same as given in Table 1. Inferences are shown from oldest (top) to most recent (bottom).

* Reference for the data being analysed.

Table 3 Sampling properties of the data sets analysed in Tables 1 and 2

Locus or DNA region and reference	Pattern of inheritance	Number of individuals	Number of locations	Number of haplotypes	TMRCA* (Myr ago)
mtDNA ^{36,37}	Maternal haploid	1,389	12	77	0.24
mtDNA ^{11,12}	Maternal haploid	532	21	109	0.24
mtDNA ¹⁰	Maternal haploid	1,178	14	127	0.24
mtDNA ⁹	Maternal haploid	53	16	51	0.24
SRY and YAP ¹³	Paternal haploid	1,544	35	10	0.23
X non-coding ¹⁵	Haplo-diploid	69	29	33	0.67
PDHA1 ¹⁴	Haplo-diploid	35	8	11	1.91
Haemoglobin β ¹⁶	Diploid	160	9	14	1.63
EDN ²⁰	Diploid	67	4	9	1.15
ECF ²⁰	Diploid	67	4	8	1.09
MC1R ¹⁸	Diploid	121	6	7	0.85
MX1 ¹⁷	Diploid	354	21	10	8.5
MS205 ¹⁹	Diploid	389	15	15	1.25

*Time to the most recent common ancestral haplotype.

rate at that locus. Even for the rapidly evolving mtDNA, populations isolated from one another for a few thousand years would not be detected as fragmented in a nested clade analysis.

The temporal sampling period also varies across loci. A haplotype tree contains information up to but excluding the final coalescent event to a common ancestral molecule. Different genes have different expected coalescent times as a function of their pattern of inheritance (Table 4). Hence, the unisexual haploid elements (mtDNA and Y-DNA) are expected to detect only relatively recent events, whereas the bisexual X-linked and autosomal genes are expected to detect far older events. There is also a large variance in coalescent times across loci sharing a common pattern of inheritance³⁶. For example, the expected coalescent time for a large sample at an autosomal locus under neutrality is $4N_{ef}$ and the variance is about $4.64N_{ef}$ (ref. 2), where N_{ef} is the long-term inbreeding effective size of the population (see Table 4). Thus, different DNA regions sharing a common pattern of inheritance are expected to show considerable variation in their temporal depths, as is indeed found in the DNA regions analysed (Table 3).

The dates in Table 3 were obtained using the method of ref. 37 that estimates the time to the most recent common ancestral haplotype (TMRCA) by calculating the ratio of the average nucleotide differences within the human sample to one-half the average nucleotide difference between chimpanzees and humans and multiplying the ratio by an estimate of the divergence time between humans and chimpanzees. Takahata *et al.*³⁷ used 5 million years for this divergence time, but recent fossil finds^{38,39} indicate that it is more likely to be at least 6 million years, a figure still compatible with the molecular data⁴⁰. Hence, the estimates³⁷ for PDHA1, Xq13.3, the β -globin locus, mtDNA and Y-DNA are recalculated with a calibration point of 6 million years. This same estimator is applied to the remaining loci used in this study, with the results shown in Table 3. Two loci require special consideration. The first is MX1, which has an estimated TMRCA of 8.5 million years ago, a date older than the chimpanzee–human divergence of 6 million years ago. A coalescent time older than the species is a true biological possibility known as trans-specific polymorphism. The TMRCA for MS205 had already been estimated at 1 million years⁴¹ with the calibration date of 5 million years, but is given as 1.2 million years in

Table 3 using the 6 million year calibration date.

Rannala and Bertorelle⁴² argue that phylogenetic dating approaches like that of ref. 37 are also appropriate for dating the ages of the older clades within a haplotype tree but are unreliable for the more recent clades. In particular, the procedure of ref. 37 ignores the sampling error in estimating the average nucleotide differences both within and between species and clades of haplotypes, and the error in the coalescent process itself due to genetic drift⁴⁰. These errors can be quite large for recent haplotypes⁴². Therefore, the procedure³⁷ is used only to date the inferences at the higher (older) nesting levels in the GEODIS analyses. The age of an inferred process or event is estimated as the age of the youngest monophyletic clade that contributed in a statistically significant fashion to the inference. This represents the time in the population's history during which all haplotype lineages affected by the event were present but more derived haplotype lineages not affected by the event had not yet evolved. This method of ageing reflects the fact that the mutations that occur during the event or process provide the strongest signal for biological inference under a nested clade analysis. For example, a mutation that occurs in a population in the act of expanding into a new geographical area provides the signal used by the nested clade analysis to infer range-expansion events^{21,23}. The age of the youngest clade marking an event or process is therefore expected to be largely coincident with the age of the event itself in most cases. However, there are exceptions. For example, a mutation occurring in a population shortly before a range expansion will have a large geographical range for its frequency once the expansion has occurred—another signal for range expansion in the nested clade analysis^{21,23}. Alternatively, a mutation occurring in the new geographical area shortly after a range expansion could create a new haplotype with a limited geographical range that is located far from the geographical centre of its ancestral haplotype—yet another signal for range expansion in the nested clade analysis^{21,23}. Thus, the estimated age of an event could sometimes be older or younger than the actual event.

Because the procedure of ref. 37 does not take into account the potentially large error associated with the coalescent process, I used my method⁴⁰ to fit a gamma distribution with a mean equal to the date estimated³⁷ and with a variance given by a neutral coalescent process given the amount of pairwise diversity observed within the monophyletic clade being dated. The ages of the inferred events and processes are therefore treated as random variables rather than constants. Because the dating of an event from any one gene could be either younger or older than the actual event and the error associated with any one gene can be large, a conservative approach is taken in reserving hard inference only for those events for which there is significant overlap (5% or more) in the probability distributions of the inferred ages from more than one gene. By requiring cross-validation over genes, a parsimonious interpretation of the number of events is obtained, and the dating of these events can be averaged over the genes to minimize errors associated

Table 4 Expected coalescent times for different inheritance patterns

DNA region	Pattern of inheritance	Expected coalescence time in generations
mtDNA	Maternal haploid	$2N_{ef} \approx N_{ef}^*$
Y-Chromosomal DNA	Paternal haploid	$2N_{efm} \approx N_{ef}$
X-linked DNA	Bisexual haplo-diploid, equal sex ratio	$3N_{ef}$
Autosomal DNA	Bisexual diploid	$4N_{ef}$

* N_{ef} is the long-term inbreeding effective size of the population, N_{efm} the inbreeding effective size of females, and N_{efm} the inbreeding effective size of males. The approximations to the expected coalescence time are made under the assumption that $N_{ef} = N_{efm} = N_{ef}/2$.

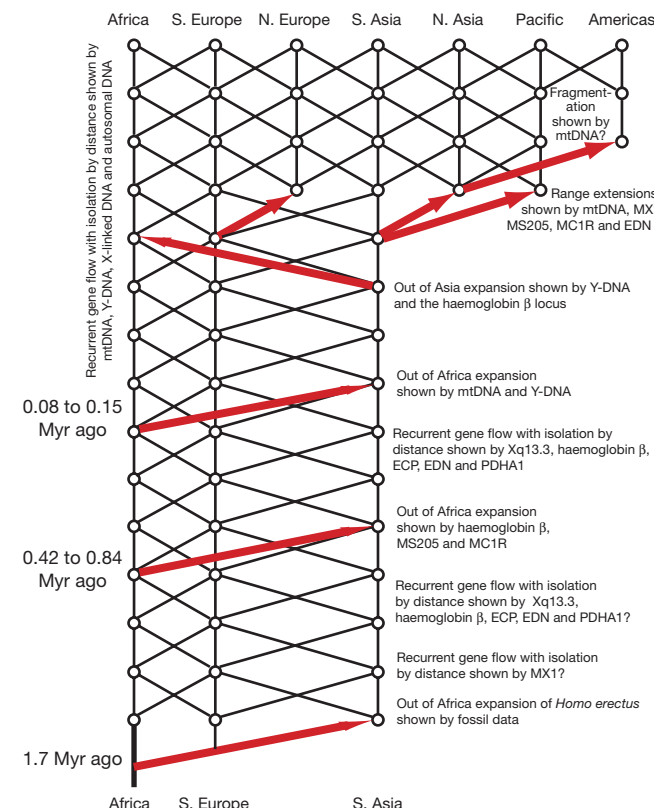


Figure 1 A new model of recent human evolution. All statistically significant inferences in Tables 1 and 2 are incorporated into this single model. Major expansions of human populations are indicated by red arrows. Genetic descent is indicated by vertical lines, and gene flow by diagonal lines. The timing of inferences lacking resolution at the 5% level and/or not validated by more than one locus are indicated by question marks.

with any one gene. Any event that is not cross-validated is regarded as tentative and questionable. There is no biological reason why dates of inferences of recurrent processes such as gene flow should be concordant across loci. However, a conservative approach is taken here as well in which those time intervals indicating gene flow that are not cross-validated at the 95% level are regarded as tentative and questionable.

Inferences on recent human evolution

All the inferences in Tables 1 and 2 can be placed into a single, internally consistent model of recent human evolution (Fig. 1). This model accepts the fossil evidence that the human lineage started in Africa and initially spread out of Africa about 1.7 million years ago⁴³. The oldest inferences for four genes (Table 2) are recurrent gene flow constrained by isolation by distance at the *MX1* locus (3.41 million years ago or Myr ago), *PDHA1* (0.86 Myr ago), *EDN* (0.69 Myr ago) and *ECP* (0.58 Myr ago). Two other genes yield the same inference in middle clades (Table 1) that date to a similar time period (β -globin, 0.49 Myr ago; and Xq 13.3, 0.48 Myr ago). Figure 2 shows the distributions for these ages. All of these six inferences of isolation by distance have broadly overlapping time intervals with the exception of *MX1*, the locus that samples the oldest time period (Table 3). The 95% confidence interval for the inference of isolation by distance from *MX1* ranges from 8.5 Myr ago to 0.61 Myr ago, indicating a low confidence in the timing of this gene flow. The spatial pattern of gene flow detected with *MX1* could not have existed until after there were human populations in Eurasia, so the fossil date of 1.7 Myr ago places a biological bound on the older limit of the age interval. This confidence interval is consistent with recurrent gene flow constrained by isolation by

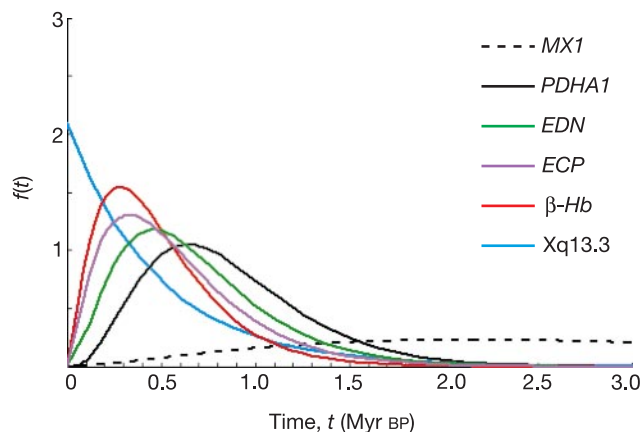


Figure 2 The distributions for the ages of the youngest clade contributing to a significant inference of gene flow constrained by isolation by distance at the highest nesting level for the genes *MX1*, *PDHA1*, *EDN* and *ECP*, and at intermediate nesting levels for the β -globin (β -Hb) locus and the Xq13.3 region. The x axis gives the time in millions of years before present (BP), and the y axis gives the gamma probability distribution, $f(t)$, that was fitted to the clade data.

distance shortly after the original expansion out of Africa because 21% of the probability mass of this gamma distribution is associated with times less than 1.7 Myr ago. However, no other inference of isolation by distance has 5% or more of its probability mass extending that far back in time. Moreover, *MX1* is an outlier in TMRCA (Table 3), which raises the possibility that natural selection has shaped the coalescent process at this locus. Therefore, the evidence for recurrent gene flow among the *Homo erectus* populations shortly after they first came out of Africa is tentative and of questionable validity.

Figure 2 shows that as time progresses, the evidence for gene flow with isolation by distance among Old World human populations is increasingly cross-validated even when *MX1* is excluded. For example, all five remaining inferences of restricted gene flow have more than 5% of their probability mass on times 1 Myr ago and older: *PDHA1* ($P = 0.32$), *EDN* ($P = 0.19$), *ECP* ($P = 0.13$), β -globin (0.08), and Xq13.3 (0.12) and the probability that one or more of these loci is detecting recurrent gene flow 1 Myr ago or older is 0.61. This probability increases to 0.95 by 0.61 Myr ago or older. These results indicate that recurrent gene flow occurred among Old World human populations from the present back to at least 600,000 years ago, and probably even further into the past.

The highest clade level inferences for five genes are range expansions out of Africa (mtDNA at 0.13 Myr ago; Y-DNA at 0.09 Myr ago; *MC1R* at 0.64 Myr ago; and β -globin at 0.82 Myr ago) and a range expansion of ambiguous origin (*MS205* at 0.63 Myr ago). Figure 3 shows the gamma distributions associated with the ages of these expansion events. The β -globin inference has 5.6% of its probability mass at 1.7 Myr ago or older, but all the other inferred range expansions have much less probability mass at 1.7 Myr ago or older. Hence, there is no cross-validated inference of marking the original expansion of *Homo erectus* out of Africa. Figure 3 shows that the five inferences fall into two broadly overlapping classes; the mtDNA and Y-DNA that are tightly clustered around an expansion event out of Africa at about 100,000 years ago, and the three autosomal loci with means between 0.64 to 0.82 Myr ago. To test the null hypothesis that all five loci are compatible with a single range expansion event, the lower age bound was found such that 0.1% of the β -globin probability distribution was above this age (this locus has the oldest inferred expansion event) and the upper age bound was found such that 0.1% of the Y-DNA probability distribution was below this age (this locus has the youngest inferred expansion event). Any age outside of this interval would be strongly

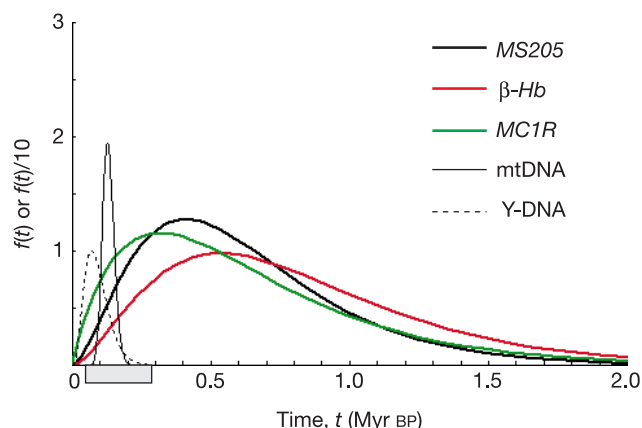


Figure 3 The distributions for the ages of the youngest clade contributing to a significant inference of a population range expansion for mtDNA, Y-DNA, *MC1R*, *MS205*, and the β -globin (β -Hb) locus. The y-axis gives the gamma probability distribution, $f(t)$, that was fitted to the clade data. Because the probability mass of the gamma distributions for mtDNA and Y-DNA are heavily concentrated into a small, recent, time interval, their probability distributions divided by ten, $f(t)/10$, are plotted instead. The grey area on the time axis indicates the interval of most likely overlap among these five distributions.

rejected (at the 0.1% level) on this basis of Y-DNA or β -globin alone. This interval spans between 0.0474 to 0.2906 Myr ago and is shown in grey in Fig. 3. The probability that all five loci are detecting an event in this age interval is 0.003. Because the expansion event detected with *MS205* is of ambiguous geographical origin, the calculation was repeated excluding that locus to focus specifically upon out-of-Africa expansion events. With this exclusion, the hypothesis of a single out-of-Africa expansion event is rejected with $P = 0.018$. The broad overlap of the mtDNA and Y-DNA distributions (Fig. 3) indicates that the null hypothesis of these two genetic elements marking the same out-of-Africa range-expansion event cannot be rejected, and similarly the null hypothesis that the three nuclear markers in Fig. 3 are marking the same expansion event cannot be rejected. The parsimonious explanation of Fig. 3 is therefore a minimum of two out-of-Africa expansion events, both of which are cross-validated by more than one locus. Combining the three nuclear genes, the 95% confidence interval for the older out-of-Africa range-expansion event is 0.42 to 0.84 Myr ago. Combining the mtDNA and Y-DNA distributions, the 95% confidence interval for the more recent out-of-Africa expansion event is 0.08 to 0.15 Myr ago.

The GEODIS analyses indicate that the most recent out-of-Africa expansion event was not a replacement event. If it had been, the three significant genetic signatures of the older expansion event (Fig. 3) and the six significant genetic signatures of older recurrent gene flow (Fig. 2) would have been wiped away. Although there is considerable error in dating any single inference from only one gene, an out-of-Africa replacement event would require that all nine significant inferences found in all eight bisexually inherited nuclear loci examined would have to be in error simultaneously. Moreover, the dating errors would have to be large in all nine cases and in the same direction. The hypothesis of a recent out-of-Africa replacement event is therefore strongly rejected. Indeed, this is the most highly cross-validated conclusion of the entire analysis because every gene region that is expected to contain information about events or processes substantially older than 0.15 Myr ago leads to a rejection of recent replacement. Hence, this is a genome-wide conclusion and not a locus-specific conclusion. The out-of-Africa expansion that took place between 0.08 to 0.15 Myr ago therefore represented a major movement of peoples characterized by interbreeding and not replacement.

It is likely that the earlier out-of-Africa expansion event between

0.42 to 0.84 Myr ago was also characterized by interbreeding and not replacement. Even excluding the *MX1* result, the other distributions shown in Fig. 2 jointly define a probability of 0.78 of recurrent gene flow before 0.84 Myr ago, a signal that would have been erased by replacement. Hence, the evidence suggests that the intermediate out-of-Africa expansion event shown in Fig. 1 was also not a replacement event, but this suggestion must be regarded as tentative because the probability of such ancient gene flow is less than 0.95.

Proceeding downwards in the nested inferences shown in Tables 1 and 2, several recent inferences are encountered of recurrent but restricted gene flow and of population expansions. Rannala and Bertorelle⁴² warned that recent events cannot be reliably dated phylogenetically. Their warning is confirmed by this analysis. The youngest clade marking many of these recent inferences is a tip haplotype that has no mutational derivatives, resulting in an average within-nucleotide difference of zero. It is impossible to use any method of phylogenetic dating in these cases. Even when the average within-nucleotide difference is greater than zero, the estimate is likely to be unreliable because the number of informative mutations for recent events is so small⁴². For example, the expansion of humans into the Americas as detected by *MS205* (Table 2) is dated phylogenetically at 449,000 years ago, with a 95% confidence interval of 1.18 Myr ago to 68,000 years ago. Even 68,000 years ago appears to be too old for this event¹⁷. The time estimate in this case depends upon just two mutational events, and as predicted⁴², the resulting estimate seems to be unreliable. Rannala and Bertorelle⁴² suggest that recent haplotypes should be dated instead through a population genetic approach that uses models of demography, mutation, and/or recombination to estimate allele age. However, the population genetic approach in turn is limited by the reliability of these underlying models. In light of these considerations, the recent inferences found in the lower hierarchical levels shown in Tables 1 and 2 will undoubtedly be more accurately dated by palaeontological and archaeological evidence than by genetic data. Therefore, no estimates of dates of these inferences will be given here.

Most of these recent inferences come from Y-DNA or mtDNA (Tables 1 and 2), and their nesting position means that they must be more recent than the most recent out-of-Africa expansion event found at the highest (oldest) nesting levels for Y-DNA and mtDNA. Many of these recent inferences are expansions of humans into areas not formerly occupied; such as northern Eurasia, the Pacific (including Australia), and the Americas. However, one of these recent expansion events did occur within the range of previous human occupation. Both the Y-DNA and β -globin locus detect a recent out-of-Asia expansion event that did not erase earlier genetic signals (Table 1) and therefore was an expansion/interbreeding event rather than an expansion/replacement event. Indeed, the failure to detect this recent out-of-Asia expansion event with mtDNA despite good sample sizes and genetic resolution may indicate that this recent out-of-Asia expansion event was primarily male-mediated, and therefore had to be characterized by interbreeding in order to leave a genetic signature.

Genetic contact among human populations after the latest out-of-Africa expansion is also indicated by several inferences at the lower clade levels of recurrent gene flow constrained by isolation by distance with occasional long-distance dispersal (Tables 1 and 2), a conclusion confirmed by other genetic analyses³². The only evidence for significant fragmentation (long-term genetic isolation) among the human populations who have left living descendants is the recent genetic isolation between Amerindians and the rest of humanity that followed the colonization of the Americas (Table 1). All of these inferences are summarized in Fig. 1.

Discussion

The model of recent human evolution shown in Fig. 1 is dominated

by genetic interchange and a special role for Africa. I consider first genetic interchange. African and Eurasian populations were linked by recurrent gene flow, certainly over the last half a million years, and probably longer. Overlaid upon this gene-flow trellis are occasional major movements out of Africa and out of Asia that enhanced gene interchange through interbreeding. More recently, population expansions acted to extend the geographical range of the human species and to establish additional areas linked by gene flow. This model emphasizes that genetic interchange among human populations, facilitated both by gene flow and range expansions coupled with interbreeding, has been a major force in shaping the human species and its spatial pattern of genetic diversity.

Second, Fig. 1 reveals the special role that African populations have played in human evolution. There were at least two major movements of peoples out of Africa after the original spread of *Homo erectus*. This inference is consistent with the archaeological record of cultural expansions out of Africa (Acheulean) in the middle Pleistocene^{44–48}. These Acheulean cultural expansions broadly overlap the time frame of the middle out-of-Africa expansion event shown in Fig. 1, indicating that this expansion involved both people and ideas coming out of Africa and interacting with local populations in Eurasia. This expansion is also compatible with the fossil data. After the initial expansion of *Homo erectus* out of Africa about 1.7 Myr ago, there was little change in average brain size up to 700,000 years ago¹. By 400,000 to 500,000 years ago, average cranial capacities had shown a substantial increase¹. The time period of this transition in cranial capacity overlaps extensively with the time period for the older out-of-Africa expansion event shown in Fig. 3.

The most recent out-of-Africa expansion event shown in Figs 1 and 3 is also compatible with fossil evidence. Many 'modern' traits (such as high, rounded skulls; small brow ridges; a vertical forehead; and a noticeable chin) first appear in Africa about 130,000 years ago, followed by an expansion out-of-Africa more than 90,000 years ago¹. This time frame overlaps extensively with the out-of-Africa expansion marked by the mtDNA and Y-DNA distributions in Fig. 3, implying that many of these traits could have been carried into Eurasia by this African population range expansion. Other traits, however, do not display any significant changes before, during or after this most recent expansion out of Africa¹. This later set of traits is difficult to reconcile with a population replacement, but is compatible with this most recent out-of-Africa expansion event being characterized by interbreeding. With interbreeding, mendelian inheritance allows some traits to spread while others do not. Moreover, living humans are still polymorphic for 'modern' traits, and the frequencies of different 'modern' traits show heterogeneity in their present geographical distributions¹. The current spatial and frequency heterogeneity in 'modern' traits undercuts the idea of a global replacement of an 'archaic' type by a 'modern' type but is consistent with a trait-based evolution of humans that is allowed under expansion with interbreeding. The model in Fig. 1 indicates the recent fossil evidence should be interpreted in terms of traits and not population types.

The genetic impacts of Africa upon the entire human species is large because of at least three major expansions out of Africa, although the genetic impact is not as complete as it would be under total replacement. This model is similar to earlier models that have emphasized the role of out-of-Africa population expansion coupled with gene flow and not replacement, such as the assimilation model of Smith *et al.*⁴⁹, the multiregional model with expansions followed by admixture of Wolpoff *et al.*⁵⁰, and the 'mostly out of Africa' model of Relethford¹.

The predicted large genetic impact of African populations explains the results of Takahata *et al.*³⁷ that about 90% of the haplotype trees in the nuclear genome appear to be rooted in Africa. These results³⁷ also falsify a total replacement hypothesis, which predicts that all haplotype trees with coalescent times greater

than 100,000 years must be rooted in Africa. All of the haplotype trees considered³⁷ have expected coalescent times greater than 100,000 years, so 100% of such old trees should have African roots under complete replacement, and not the observed 90%.

The results given here show the importance of examining many DNA regions with a common analytical technique in making phylogeographic inferences. Indeed, the clearest result from Tables 1 and 2 is how incomplete our view of human evolution would be if it were based upon just one locus or DNA region. As more DNA regions are examined, additional insights into human evolution are sure to follow. However, this current analysis already demonstrates the inadequacies of both the out-of-Africa replacement model and of a simple trellis model. Humans expanded again and again out of Africa, but these expansions resulted in interbreeding, not replacement, and thereby strengthened the genetic ties between human populations throughout the world. □

Received 30 August 2001; accepted 3 January 2002.

1. Relethford, J. H. *Genetics and the Search for Modern Human Origins* (Wiley, New York, 2001).
2. Stoneking, M. & Soodyall, H. Human evolution and the mitochondrial genome. *Curr. Opin. Genet. Dev.* **6**, 731–736 (1996).
3. Wolpoff, M. H., Hawks, J. & Caspari, R. Multiregional, not multiple origins. *Am. J. Phys. Anthropol.* **112**, 129–136 (2000).
4. Weidenreich, F. *Apes, Giants, and Man* (Univ. Chicago Press, Chicago, 1946).
5. Cann, R. L., Stoneking, M. & Wilson, A. C. Mitochondrial DNA and human evolution. *Nature* **325**, 31–36 (1987).
6. Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K. & Wilson, A. C. African populations and the evolution of human mitochondrial DNA. *Science* **253**, 1503–1507 (1991).
7. Maddison, D. R. African origin of human mitochondrial DNA reexamined. *Syst. Zool.* **40**, 355–363 (1991).
8. Templeton, A. R. Human origins and analysis of mitochondrial DNA sequences. *Science* **255**, 737 (1992).
9. Ingman, M., Kaessmann, H., Pääbo, S. & Gyllenstein, U. Mitochondrial genome variation and the origin of modern humans. *Nature* **408**, 708–713 (2000).
10. Sykes, B., Leifoff, A., Low-Beer, J., Tetzner, S. & Richards, M. The origins of the Polynesians: an interpretation from mitochondrial lineage analysis. *Am. J. Hum. Genet.* **57**, 1463–1475 (1995).
11. Torroni, A. *et al.* Asian affinities and continental radiation of the four founding native American mtDNAs. *Am. J. Hum. Genet.* **53**, 563–590 (1993).
12. Torroni, A. *et al.* mtDNA variation of aboriginal Siberians reveals distinct genetic affinities with Native Americans. *Am. J. Hum. Genet.* **53**, 591–608 (1993).
13. Hammer, M. F. *et al.* Out of Africa and back again: Nested cladistic analysis of human Y chromosome variation. *Mol. Biol. Evol.* **15**, 427–441 (1998).
14. Harris, E. E. & Hey, J. X. Chromosome evidence for ancient human histories. *Proc. Natl Acad. Sci. USA* **96**, 3320–3324 (1999).
15. Kaessman, H., Heifßig, F., Haeseler, A. V. & Pääbo, S. DNA sequence variation in a non-coding region of low recombination on the human X chromosome. *Nature Genet.* **22**, 78–81 (1999).
16. Harding, R. M. *et al.* Archaic African and Asian lineages in the genetic ancestry of modern humans. *Am. J. Hum. Genet.* **60**, 772–789 (1997).
17. Jin, L. *et al.* Distribution of haplotypes from a chromosome 21 region distinguishes multiple prehistoric human migrations. *Proc. Natl Acad. Sci. USA* **96**, 3796–3800 (1999).
18. Rana, B. K. *et al.* High polymorphism at the human melanocortin 1 receptor locus. *Genetics* **151**, 1547–1557 (1999).
19. Rogers, E. J., Shone, A. C., Alonso, S., May, C. A. & Armour, J. A. L. Integrated analysis of sequence evolution and population history using hypervariable compound haplotypes. *Hum. Mol. Genet.* **9**, 2675–2681 (2000).
20. Zhang, J. & Rosenberg, H. F. Sequence variation at two eosinophil-associated ribonuclease loci in humans. *Genetics* **156**, 1949–1958 (2000).
21. Templeton, A. R., Routman, E. & Phillips, C. Separating population structure from population history: a cladistic analysis of the geographical distribution of mitochondrial DNA haplotypes in the Tiger Salamander, *Ambystoma tigrinum*. *Genetics* **140**, 767–782 (1995).
22. Posada, D., Crandall, K. A. & Templeton, A. R. GeoDis: a program for the cladistic nested analysis of the geographical distribution of genetic haplotypes. *Mol. Ecol.* **9**, 487–488 (2000).
23. Templeton, A. R. Nested clade analyses of phylogeographic data: testing hypotheses about gene flow and population history. *Mol. Ecol.* **7**, 381–397 (1998).
24. Bernatchez, L. The evolutionary history of brown trout (*Salmo trutta* L.) inferred from phylogeographic, nested clade, and mismatch analyses of mitochondrial DNA variation. *Evolution* **55**, 351–379 (2001).
25. Gomez, A., Carvalho, G. R. & Hunt, D. H. Phylogeography and regional endemism of a passively dispersing zooplankton: mitochondrial DNA variation in rotifer resting egg banks. *Proc. R. Soc. Lond. B* **267**, 2189–2197 (2000).
26. Nielson, M., Lohman, K. & Sullivan, J. Phylogeography of the tailed frog (*Ascaphus truei*): Implications for the biogeography of the Pacific Northwest. *Evol.* **55**, 147–160 (2001).
27. Turner, T. F., Trexler, J. C., Harris, J. L. & Haynes, J. L. Nested cladistic analysis indicates population fragmentation shapes genetic diversity in a freshwater mussel. *Genetics* **154**, 777–785 (2000).
28. Templeton, A. R., Crandall, K. A. & Sing, C. F. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. *Genetics* **132**, 619–633 (1992).
29. Posada, D. TCS 1.06 (Provo, Utah, 2000); available at http://bioag.byu.edu/zoology/crandall_lab/tcs.htm.
30. Crandall, K. A. & Templeton, A. R. Empirical tests of some predictions from coalescent theory with applications to intraspecific phylogeny reconstruction. *Genetics* **134**, 959–969 (1993).

31. Templeton, A. R. in *Conceptual Issues in Modern Human Origins Research* (eds Clark, G. A. & Willermet, C. M.) 329–360 (Aldine de Gruyter, New York, 1997).
 32. Templeton, A. R. Human races: A genetic and evolutionary perspective. *Am. Anthropol.* **100**, 632–650 (1998).
 33. Templeton, A. R. Using phylogeographic analyses of gene trees to test species status and processes. *Mol. Ecol.* **10**, 779–791 (2001).
 34. Excoffier, L. & Langaney, A. Origin and differentiation of human mitochondrial DNA. *Am. J. Human Genet.* **44**, 73–85 (1989).
 35. Excoffier, L. Evolution of human mitochondrial DNA: evidence for departure from a pure neutral model of populations at equilibrium. *J. Mol. Evol.* **30**, 125–139 (1990).
 36. Donnelly, P. & Tavaré, S. Coalescents and genealogical structure under neutrality. *Annu. Rev. Genet.* **29**, 401–421 (1995).
 37. Takahata, N., Lee, S.-H. & Satta, Y. Testing multiregionality of modern human origins. *Mol. Biol. Evol.* **18**, 172–183 (2001).
 38. Haile-Selassie, Y. Late Miocene hominids from the Middle Awash, Ethiopia. *Nature* **412**, 178–181 (2001).
 39. Pickford, M. & Senut, B. The geological and faunal context of the Late Miocene hominid remains from Lukeino, Kenya. *C.R. Acad. Sci. IIA* **332**, 145–152 (2001).
 40. Templeton, A. R. The “Eve” hypothesis: a genetic critique and reanalysis. *Am. Anthropol.* **95**, 51–72 (1993).
 41. Alonso, S. & Armour, J. A. L. A highly variable segment of human subterminal 16p reveals a history of population growth for modern humans outside Africa. *Proc. Natl. Acad. Sci. USA* **98**, 864–869 (2001).
 42. Rannala, B. & Bertorelle, G. Using linked markers to infer the age of a mutation. *Hum. Mutat.* **18**, 87–100 (2001).
 43. Gabunia, L. *et al.* Earliest Pleistocene hominid cranial remains from Dmanisi, Republic of Georgia: Taxonomy, geological setting, and age. *Science* **288**, 1019–1025 (2000).
 44. Aguirre, E. & Carbonell, E. Early human expansions into Eurasia: The Atapuerca evidence. *Quat. Int.* **75**, 11–18 (2001).
 45. Bar-Yosef, O. & Belfer-Cohen, A. From Africa to Eurasia—early dispersals. *Quat. Int.* **75**, 19–28 (2001).
 46. Hou, Y. M. *et al.* Mid-Pleistocene Acheulean-like stone technology of the Bose basin, South China. *Science* **287**, 1622–1626 (2000).
 47. Saragusti, I. & Goren-Inbar, N. The biface assemblage from Gesher Benot Ya’aqov, Israel: illuminating patterns in “Out of Africa” dispersal. *Quat. Int.* **75**, 85–89 (2001).
 48. Otte, M. in *Archaeogenetics: DNA and the Population Prehistory of Europe* (eds Renfrew, C. & Boyle, K.) 41–44 (Univ. Cambridge, Cambridge, 1999).
 49. Smith, F. H., Falsetti, A. B. & Donnelly, S. M. Modern human origins. *Yb. Physical Anthropol.* **32**, 35–68 (1989).
 50. Wolpoff, M., Thorne, A. G., Smith, F. H., Frayer, D. W. & Pope, G. G. in *Origins of Anatomically Modern Humans* (eds Nitecki, M. H. & Nitecki, D. V.) 175–200 (Plenum, New York, 1994).
- Supplementary Information** accompanies the paper on *Nature’s* website (<http://www.nature.com>).
- Acknowledgements**
- I thank J. Brisson, J. Hess, R. Koch, M. Kramer, R. Robertson and J. Strasburg for suggestions on an earlier draft of this manuscript. I also thank E. Trinkhaus and J. Relethford for their reviews. This work was supported in part by a Burroughs Wellcome Fund Innovation Award in Functional Genomics.
- Correspondence and requests for materials should be addressed to A.R.T. (e-mail: temple_a@biology.wustl.edu).

Identification of a cold receptor reveals a general role for TRP channels in thermosensation

David D. McKemy^{*†}, Werner M. Neuhausser^{*†} & David Julius^{*}

^{*} Department of Cellular and Molecular Pharmacology, University of California, San Francisco, California 94143-0450, USA

[†] These authors contributed equally to this work

The cellular and molecular mechanisms that enable us to sense cold are not well understood. Insights into this process have come from the use of pharmacological agents, such as menthol, that elicit a cooling sensation. Here we have characterized and cloned a menthol receptor from trigeminal sensory neurons that is also activated by thermal stimuli in the cool to cold range. This cold- and menthol-sensitive receptor, CMR1, is a member of the TRP family of excitatory ion channels, and we propose that it functions as a transducer of cold stimuli in the somatosensory system. These findings, together with our previous identification of the heat-sensitive channels VR1 and VRL-1, demonstrate that TRP channels detect temperatures over a wide range and are the principal sensors of thermal stimuli in the mammalian peripheral nervous system.

The somatosensory system detects changes in ambient temperature over a remarkably wide range. This process is initiated when a hot or cold stimulus excites sensory nerve fibres that project from dorsal root or trigeminal ganglia, which innervate regions of the trunk and head, respectively. These primary afferent neurons convert thermal stimuli into electrochemical signals (that is, action potentials) and relay sensory information to integrative centres in the spinal cord and brain^{1,2}. Noxious (painful) heat is detected by sensory neurons that respond with a 'moderate' thermal threshold of about 43 °C or with a 'high' threshold of about 52 °C (refs 3–5). Insights into the mechanisms of heat sensation have come from molecular cloning of the vanilloid receptor (VR1), an excitatory ion channel on sensory neurons that is activated by temperatures exceeding 43 °C, and by capsaicin, the main pungent ingredient in 'hot' chili peppers⁶. Electrophysiological, anatomical and genetic studies show that VR1 contributes to heat sensitivity of moderate-threshold neurons and is essential for the development of thermal hypersensitivity after tissue injury^{7,8}. A related ion channel, VRL-1, does not respond to capsaicin, but is activated by temperatures exceeding 50 °C, suggesting that it contributes to heat sensitivity of high-threshold sensory neurons⁹. Both VR1 and VRL-1 belong to the transient receptor potential (TRP) family of ion channels, and we have hypothesized that molecules of this type contribute to thermosensation over a wide temperature range⁹.

In contrast to our understanding of noxious heat sensation, little is known about how we detect cold. In mammals, cool sensation is generally believed to be mediated by a small sub-population of unmyelinated C and thinly myelinated Aδ primary afferent fibres that discharge in the innocuous temperature range (15–30 °C)^{10,11}. Responses to noxious cold (<15 °C) are also observed in these fibre types, with prevalence ranging from 10% to 100% depending on stimulus intensity and species examined^{7,12–15}. This wide variability in the literature may reflect the fact that thermal thresholds for cold-sensitive fibres are not as well defined as they are for heat, and thus fibre types that transduce sensations of innocuous cool versus noxious cold are not as firmly established. Calcium-imaging and patch-clamp studies of dissociated dorsal root ganglion (DRG) neurons have shown that cold (~20 °C) promotes calcium influx, possibly through the direct opening of calcium-permeable ion channels^{16,17}. However, several other mechanisms have been proposed to explain cold-evoked membrane depolarization, including

inhibition of background K⁺ channels¹⁸, activation of Na⁺-selective degenerin channels¹⁹, inhibition of Na⁺/K⁺ ATPases²⁰, and differential effects of cold on voltage-gated Na⁺ and K⁺ conductances²¹. Thus it is not clear whether cold excites sensory neurons by activating a discrete 'cold receptor', or by modulating a constellation of excitatory and inhibitory channels on these cells.

To clarify these issues, we looked for a cold transducer using an expression cloning approach. Our strategy was based on the paradigm presented by the vanilloid receptor, namely, to investigate how plant products such as menthol elicit a cool sensation. Fifty years ago, Hensel and Zotterman¹⁰ showed that menthol potentiates responses of trigeminal fibres to cold by shifting their thermal activation thresholds to warmer temperatures. Moreover, they proposed that cooling compounds mediate their psychophysical effects by interacting with a protein that is specifically involved in cold transduction. Subsequent studies suggested that menthol depolarizes sensory neurons by inhibiting voltage-dependent Ca²⁺ channels²² (thereby decreasing activation of Ca²⁺-dependent K⁺ channels)²³, or by directly activating calcium-permeable channels^{16,24}. However, there is currently no direct pharmacological or biochemical evidence to support the existence of a bona fide menthol-binding site on sensory neurons, nor is it clear whether menthol and cold act through the same molecular entity.

Here we show that the molecular site of menthol action is an excitatory ion channel expressed by small-diameter neurons in trigeminal and dorsal root ganglia. The cloned channel is also activated by cold (8–28 °C), proving that menthol elicits a sensation of cool by serving as a chemical agonist of a thermally responsive receptor. This cold- and menthol-sensitive receptor (CMR1) is a member of the long TRP (or TRPM) channel subfamily, making it a close molecular cousin of the heat-activated channels VR1 and VRL-1. Thus we conclude that TRP channels are the primary molecular transducers of thermal stimuli within the mammalian somatosensory system.

Activation of a Ca²⁺-permeable channel by menthol and cold

Body surfaces innervated by trigeminal fibres, such as the eye and tongue, are particularly sensitive to cold and cooling compounds²⁵. We therefore used calcium imaging and electrophysiological methods to examine responses of dissociated rat trigeminal neurons to menthol and cold. Indeed, these stimuli produced robust

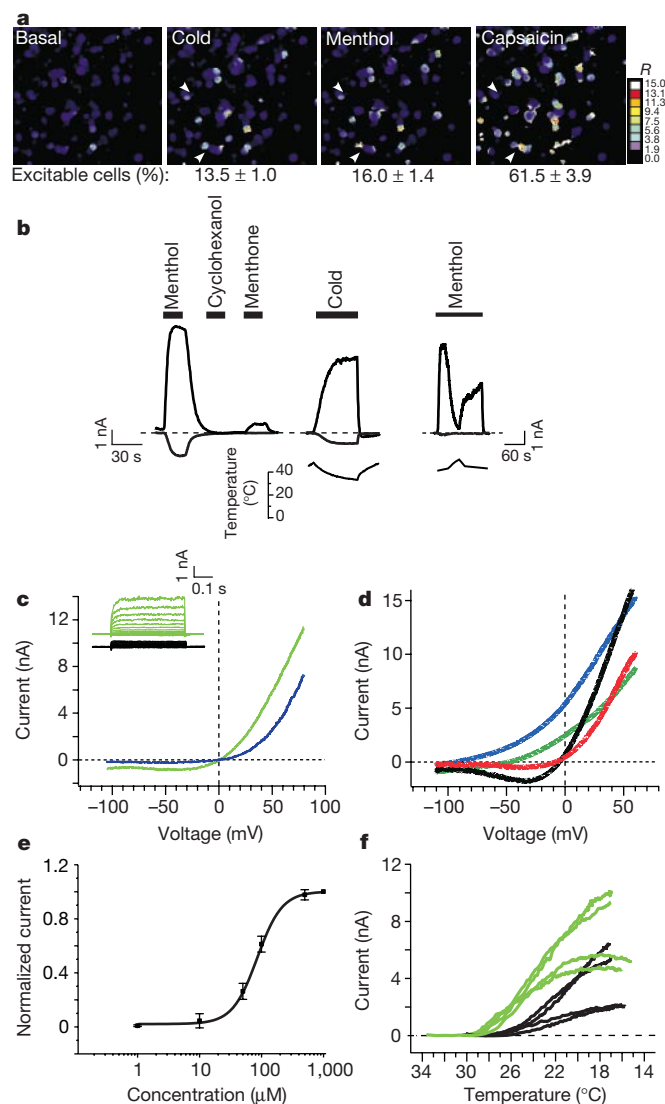


Figure 1 A subset of trigeminal neurons express an outwardly rectifying Ca^{2+} -permeable channel activated by menthol and cold. **a**, Responses of dissociated trigeminal neurons to cold (7°C), menthol ($500\ \mu\text{M}$) and capsaicin ($1\ \mu\text{M}$) were assessed by calcium imaging. Arrowheads mark menthol-responding cells that were insensitive to capsaicin. Relative calcium concentrations are indicated by the Fura-2 ratio (R). The percentage ($\pm$ s.e.m.) of excitable (potassium-sensitive) cells responding to each stimulus is indicated below. **b**, Electrophysiological responses of trigeminal neurons to menthol ($50\ \mu\text{M}$), cyclohexanol ($100\ \mu\text{M}$), menthone ($100\ \mu\text{M}$) or cold (16°C) were measured in both inward and outward directions ($V_{\text{hold}} = -60$ and $+80\ \text{mV}$, respectively). Increasing temperature of perfusate (from room temperature to 30°C) completely antagonized currents evoked by $100\ \mu\text{M}$ menthol (right). Perfusate temperatures are indicated below each current trace. **c**, Responses evoked by menthol ($50\ \mu\text{M}$, green) and cold (16°C , blue) show strong outward rectification. Inset, baseline currents (black) and menthol-evoked responses (green) in nominally Ca^{2+} -free bath solution at room temperature. **d**, Voltage ramps (-120 to $+80\ \text{mV}$ in $200\ \text{ms}$) were used to establish current-voltage relationships in different extracellular solutions. Composition of the bath and electrode solutions and calculated reversal potentials are detailed in Supplementary Information. **e**, Concentration-response curve for menthol-evoked inward currents ($V_{\text{hold}} = -60\ \text{mV}$) in trigeminal neurons. Membrane currents in each neuron were normalized to $1\ \text{mM}$ menthol at room temperature. Each point represents mean value ($\pm$ s.e.m.) from six independent neurons and were fit with the Hill equation. **f**, Temperature-response curves (from 33 to 16°C) were determined for trigeminal neurons in the presence (green) or absence (black) of $10\ \mu\text{M}$ menthol. Menthol potentiated the size of cold-evoked currents and shifted thermal thresholds from $27.1 \pm 0.5^\circ\text{C}$ to $29.6 \pm 0.3^\circ\text{C}$ ($n = 4$).

increases in intracellular free calcium in a relatively small sub-population of trigeminal neurons (Fig. 1a), consistent with work from others using DRG cultures^{16,17,24}. Menthol and cold excited a largely overlapping subset of neurons, a significant fraction of which ($54.5\% \pm 6.1$, mean $\pm$ s.e.m.) were also activated by capsaicin (Fig. 1a). Sensitivity to capsaicin is considered a functional hallmark of nociceptors (primary sensory neurons that detect noxious stimuli) and thus about half of the menthol/cold-sensitive cells may be categorized as such.

Whole-cell patch-clamp recordings from trigeminal neurons showed that menthol or cold elicited rapidly developing membrane currents (Fig. 1b) that were characterized by pronounced outward rectification (that is, responses at positive holding potentials were substantially greater than those at negative voltages) (Fig. 1c). These currents reversed close to $0\ \text{mV}$ ($E_{\text{rev}} = -3.6 \pm 1.9\ \text{mV}$ and -0.8 ± 0.3 , respectively; $n = 5$), suggesting that they result from the opening of non-selective cation channels, consistent with recent observations in DRG neurons¹⁶. Indeed, ion substitution experiments showed little discrimination among monovalent cations, but revealed significantly higher permeability P for calcium ions ($P_{\text{Ca}}/P_{\text{Na}} = 3.2$; $P_{\text{K}}/P_{\text{Na}} = 1.1$; $P_{\text{Cs}}/P_{\text{K}} = 1.2$; $n = 6$) (Fig. 1d). We found these biophysical properties particularly interesting because they are reminiscent of VR1 and other TRP channels²⁶.

Room-temperature menthol evoked responses in a dose-dependent manner (Fig. 1e) with a half-maximal effective concentration (EC_{50}) of $80 \pm 2.4\ \mu\text{M}$, a potency similar to that determined for DRG neurons using calcium imaging²⁴. Fitting these data with the Hill equation suggests that receptor activation requires the binding of more than one menthol molecule ($\eta = 2.2$). In addition to menthol, the mint plant synthesizes structural analogues that also elicit a cooling sensation, although with reduced potency. One of these, menthone, elicited very small currents in trigeminal neurons, and cyclohexanol, an inactive synthetic menthol analogue, had no effect (Fig. 1b). Cold also elicited membrane currents in a dose-dependent manner, with a characteristic temperature threshold of $27.1 \pm 0.5^\circ\text{C}$ ($n = 4$) (Fig. 1f). As reported for DRG neurons, menthol potentiated cold responses and shifted the thermal threshold to higher temperatures ($29.6 \pm 0.3^\circ\text{C}$ at $10\ \mu\text{M}$ menthol). Conversely, increasing the temperature of the perfusate (from room temperature to 30°C) completely antagonized currents evoked by $100\ \mu\text{M}$ menthol (Fig. 1b). Taken together, our findings and those of others demonstrate that menthol and cold activate a calcium-permeable channel on trigeminal and DRG sensory neurons. Moreover, our electrophysiological data show that these stimuli activate currents with very similar biophysical properties, supporting the idea of a common molecular site of action.

Expression cloning of a receptor for cooling compounds

Because menthol and cold activate native conductances with intrinsic and significant permeability to calcium ions, we reasoned that a screening strategy based on calcium imaging⁶ could be used to isolate a functional complementary DNA coding for a menthol- or cold-sensitive receptor. Because these responses are more prevalent in trigeminal cultures (14.8% versus 7.4% for DRG, $n = 745$ and $1,425$ cells, respectively), we constructed a cDNA expression library from this tissue. Pools containing about 10,000 clones were transfected into HEK293 cells (derived from human embryonic kidney). We subsequently loaded the cells with the calcium-sensitive fluorescent dye Fura-2, and microscopically examined them for changes in intracellular calcium on exposure to room-temperature menthol ($500\ \mu\text{M}$). In this way, we identified a single cDNA that conferred menthol sensitivity to transfected cells.

As noted above, menthol is one of several naturally occurring or synthetic cooling compounds that together span a wide range of relative potencies in psychophysical assays²⁵. To assess the effects of the compounds on the newly identified receptor, we expressed the cloned cDNA in *Xenopus* oocytes and measured evoked responses

using whole-cell voltage-clamp methods. Clearly, the most robust responses were elicited by the super-cooling agent icilin (AG-3-5)²⁷, which showed about 2.5-fold greater efficacy and nearly 200-fold greater potency than menthol ($EC_{50} = 0.36 \pm 0.03 \mu\text{M}$ and $66.7 \pm 3.3 \mu\text{M}$, respectively) (Fig. 2a, b). Eucalyptol, the main constituent of oil of *Eucalyptus*, also elicited membrane currents, but with lower efficacy and potency ($3.4 \pm 0.4 \text{ mM}$) than menthol or icilin. Menthone, camphor and cyclohexanol had little or no effect, even when applied at concentrations approaching their limits of solubility in aqueous buffers ($>500 \mu\text{M}$). Finally, the vanilloid-receptor agonist, capsaicin, did not elicit responses in these cells.

A more detailed biophysical analysis of the cloned receptor was carried out in transfected HEK293 cells, where menthol or icilin produced currents with nearly time-independent kinetics and steep outward rectification (Fig. 3a). Like native menthol-evoked responses in trigeminal neurons, these currents showed relatively high permeability to calcium and little selectivity among monovalent cations ($P_{\text{Ca}}/P_{\text{Na}} = 3.3$; $P_{\text{K}}/P_{\text{Na}} = 1.2$; $P_{\text{Cs}}/P_{\text{K}} = 1.1$; $n = 4-9$) (Fig. 3b). Menthol-evoked currents also showed significant desensitization ($53.9\% \pm 1.7$ decrease in peak current between the first and second application, $n = 3$), but little desensitization ($9.1\% \pm 7$,

$n = 3$) was observed in nominally calcium-free bath solution (Fig. 3c). Icilin showed even stronger desensitization, but unlike menthol this agonist was essentially inactive in the absence of extracellular calcium. Similar observations were obtained in oocytes (data not shown). When measured at the single-channel level, menthol-evoked currents also showed pronounced outward rectification. These responses were characterized by brief, transient openings and had a slope conductance of $83 \pm 3 \text{ pS}$ at positive potentials (Fig. 3d, e). We also observed events with smaller unitary currents, which may represent subconductance states of the channel or openings that were simply too brief to be resolved in our analysis.

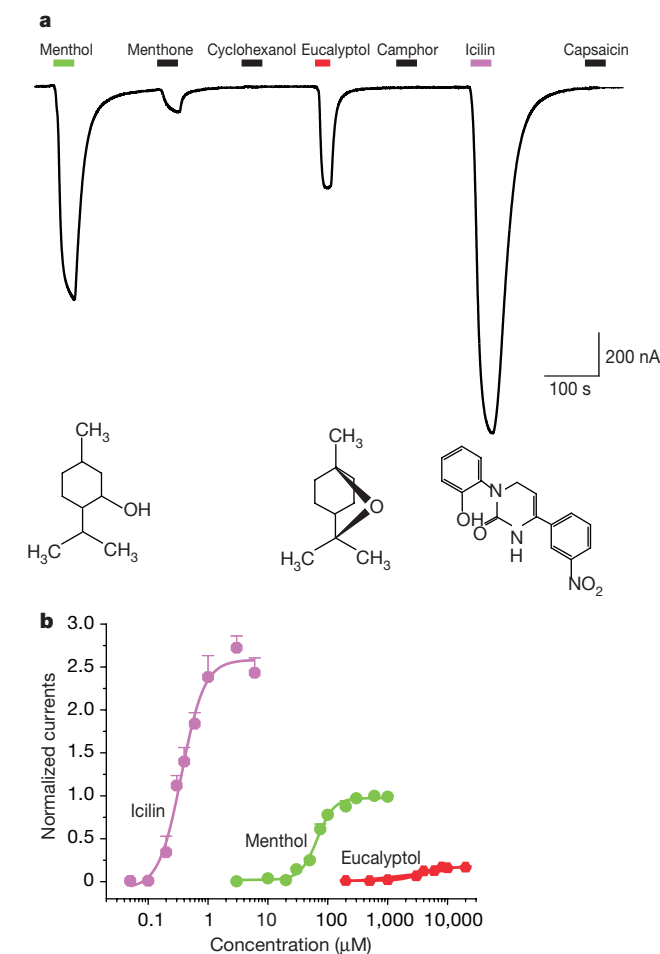


Figure 2 Cooling compounds activate the cloned receptor. **a**, An oocyte expressing the cloned receptor was exposed to consecutive applications of menthol ($100 \mu\text{M}$), menthone ($500 \mu\text{M}$), cyclohexanol ($500 \mu\text{M}$), eucalyptol (20 mM), camphor (1 mM), icilin (300 nM) and capsaicin ($1 \mu\text{M}$). Resulting membrane currents were measured under voltage clamp at -60 mV . Bars denote the duration of agonist application. Chemical structures for menthol, eucalyptol and icilin are shown below their respective responses. **b**, Concentration-response curves for icilin (pink), menthol (green) and eucalyptol (red). Responses were normalized to those evoked by $500 \mu\text{M}$ menthol. Each point represents mean values ($\pm \text{s.e.m.}$) from 4–9 independent oocytes.

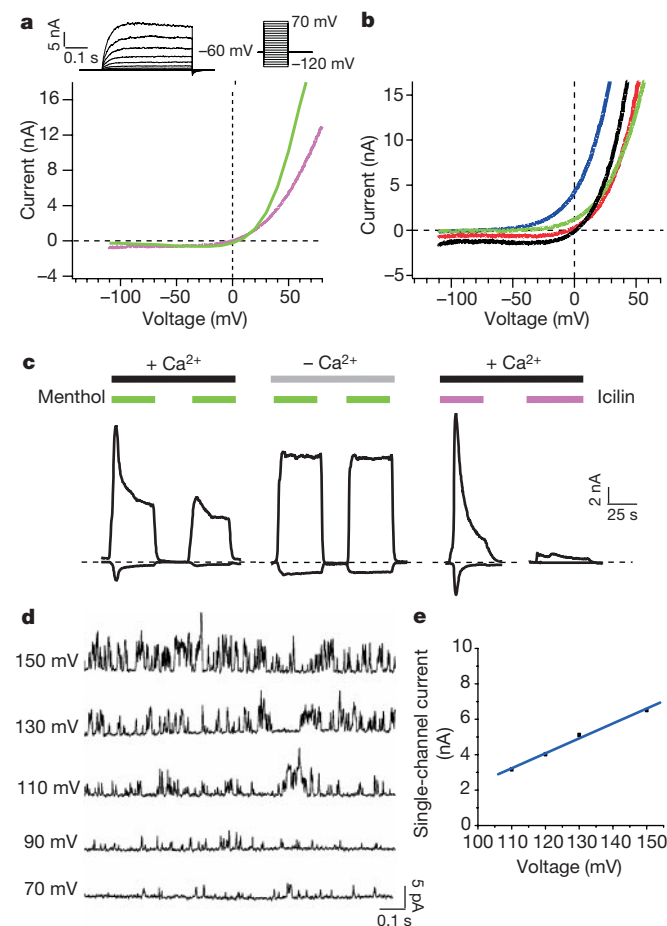


Figure 3 Electrophysiological properties of menthol-induced currents in transfected HEK293 cells. **a**, Time dependence of menthol-induced whole-cell currents were analysed using 400-ms voltage step pulses ranging from -120 to $+70 \text{ mV}$ in 10-mV steps (top right). Traces show current response induced by menthol ($50 \mu\text{M}$) at room temperature in nominally Ca^{2+} -free bath solution (top left). The current–voltage relationship (bottom) was obtained from the same pulse protocol using $200 \mu\text{M}$ menthol (room temperature) by plotting the time-independent current component as a function of membrane voltage. Menthol currents reversed at $-4.5 \pm 1.1 \text{ mV}$ ($\pm \text{s.e.m.}$, $n = 4$) and show strong outward rectification (green). A similar current–voltage relationship was obtained for $2 \mu\text{M}$ icilin (pink) from a 200-ms voltage ramp (-120 to $+80 \text{ mV}$). **b**, Voltage ramps (200 ms) ranging from -120 to $+80 \text{ mV}$ were used to record current–voltage curves in different extracellular solutions. Composition of the bath and electrode solutions and calculated reversal potentials are detailed in Supplementary Information. **c**, Current desensitization evoked by menthol ($50 \mu\text{M}$) and icilin ($2 \mu\text{M}$) is dependent on calcium. **d**, Single-channel traces were recorded from transfected HEK293 cells in the cell-attached patch configuration at different pipette potentials. **e**, Slope conductance of the amplitudes of single-channel currents was obtained by linear regression, yielding a single-channel conductance of $83 \pm 3 \text{ pS}$ ($n = 3$).

Activation of the menthol receptor by cold

To determine whether the menthol receptor is also a cold sensor, we tested its thermal responsiveness in oocytes by lowering the temperature of the perfusate from about 35 °C to about 5 °C. This elicited a robust and rapidly activating inward current (at negative holding potentials) that was remarkably consistent since the rate of temperature change ($0.2\text{--}1\text{ }^{\circ}\text{C s}^{-1}$) did not influence threshold or saturation temperature (Fig. 4a, b). Moreover, cold-evoked currents were directly proportional to temperature regardless of the direction of the temperature change (not shown). Cold-activated currents had a thermal threshold of $25.8 \pm 0.4\text{ }^{\circ}\text{C}$ and saturated at $8.2 \pm 0.3\text{ }^{\circ}\text{C}$ ($n = 12$) (Fig. 4c). Consistent with the behaviour of native cold currents, addition of a sub-activating concentration of menthol ($20\text{ }\mu\text{M}$) to the perfusate increased threshold and saturation temperature to $29.7 \pm 0.3\text{ }^{\circ}\text{C}$ and $15.6 \pm 0.4\text{ }^{\circ}\text{C}$, respectively ($n = 7$) (Fig. 4c). We found that menthol is a more efficacious agonist than cold because saturating cold-evoked currents were smaller than those obtained with a maximal dose of room-temperature menthol ($67.4\% \pm 1.9$, $n = 7$) (Fig. 4d).

We also examined cold-evoked currents in HEK293 cells expressing the menthol receptor. As observed for native cold responses (Fig. 1c), current–voltage relationships for the cloned channel showed steep outward rectification (Fig. 4e) and were markedly potentiated by a sub-activating dose of menthol ($10\text{ }\mu\text{M}$). Menthol increased cold-evoked currents in both the outward and inward direction, but the effect on the inward component was more pronounced, reminiscent of the effect of capsaicin on VR1 (ref. 28). Native cold-evoked responses in sensory fibres or cultured DRG neurons show adaptation to a prolonged thermal stimulus lasting several minutes^{16,29}. We found that receptor-transfected cells showed small, outwardly rectifying basal currents at room temperature ($\sim 22\text{ }^{\circ}\text{C}$), but responses to a subsequent $22\text{ }^{\circ}\text{C}$ stimulus were markedly larger after the cell had first been warmed to $31\text{ }^{\circ}\text{C}$ (Fig. 4f). This observation suggests that the cloned receptor also adapts to thermal challenges, and that this effect is reversed on heating. Desensitization to cold differed from that observed with menthol because it was independent of extracellular calcium (data not shown). VR1 shows similar behaviour in that desensitization to chemical (capsaicin) or thermal (heat) stimuli are dependent or independent, respectively, of calcium. Taken together, our observations show that the cloned receptor, which we now designate cold–menthol receptor type 1 (CMR1), has properties identical to endogenous cold/menthol currents observed in sensory neurons^{16,17,24}.

CMR1, member of the TRP channel family

The CMR1 cDNA contains an open reading frame of 3312 base pairs (bp) that is predicted to code for a protein of 1,104 amino acids with a relative molecular mass of 128,000 (M_r 128 K) (Fig. 5a). Database searches revealed significant homology between this deduced sequence and members of the TRP ion channel family, particularly with the subgroup of long TRP (or TRPM) channels, so named for their characteristically large amino- and carboxy-terminal cytoplasmic tails³⁰. Among members of this subfamily, TRPM2 and TRPM7 have been electrophysiologically characterized and shown to behave as bifunctional proteins in which enzymatic activities associated with their long C-terminal domains are believed to regulate channel opening^{30–32}. In contrast, CMR1 has a significantly shorter C-terminal region (Fig. 5b) and does not contain any obvious enzymatic domains that might be associated with channel regulation.

Further search of the literature showed that CMR1 is 92% identical to a human sequence originally identified as a prostate-specific transcript³³. This presumptive TRP channel, called trp-p8 (or TRPM8), is thus likely to be the human orthologue of rat CMR1. Aside from prostate epithelium, TRPM8 was found to be expressed

by a variety of tumours, including prostate, melanoma, colorectal and breast carcinoma³³. The presence of this channel in sensory neurons was not assessed and we therefore carried out northern blot and *in situ* hybridization studies to examine expression of CMR1 in rat trigeminal and dorsal root ganglia. Indeed, transcripts of about 6 and 4.5 kilobases (kb) were detected in both neuronal tissues (Fig. 6a), similar to that reported for TRPM8 expression in human prostate³³. At the cellular level, CMR1 transcripts were found in a subset of sensory neurons with small-diameter cell bodies ($18.2 \pm 1.1\text{ }\mu\text{m}$ and $21.6 \pm 0.5\text{ }\mu\text{m}$ in dorsal root and trigeminal ganglia, respectively) (Fig. 6b), similar in size to VR1-expressing cells ($19.2 \pm 0.3\text{ }\mu\text{m}$)⁹. CMR1 transcripts were more prevalent in trigeminal than dorsal root ganglia and were conspicuously absent from the vast majority of larger-diameter cells, consistent with our physiological observations using neuronal cultures. These findings suggest that CMR1 is expressed by a sub-population of C fibres (and possibly A δ fibers) representing less than 20% of all primary afferent neurons.

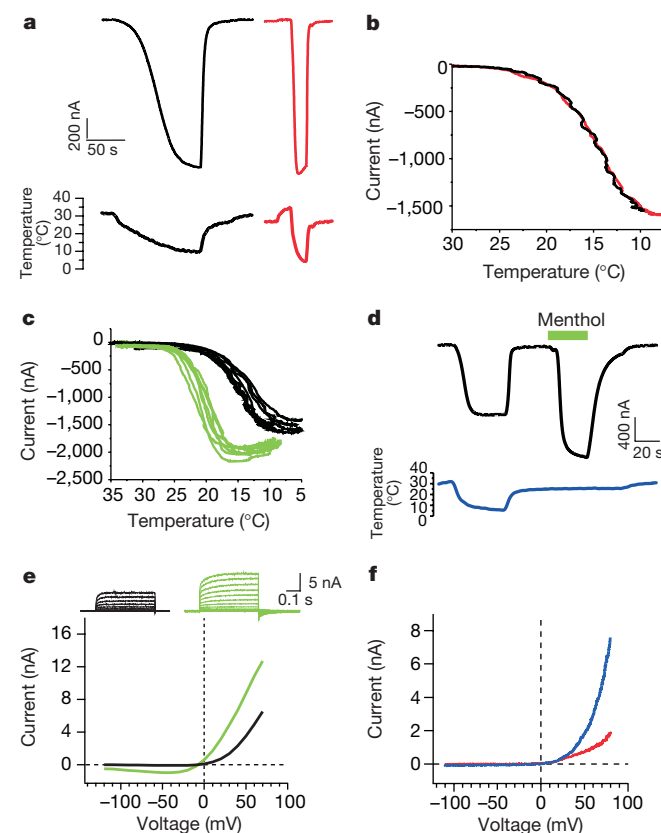


Figure 4 The menthol receptor is sensitive to cold. **a**, Inward currents (top) were evoked in the same menthol receptor-expressing oocyte by repetitive decreases in perfusate temperature. Cooling ramps (bottom) were applied at two different rates ($0.2\text{ }^{\circ}\text{C s}^{-1}$, black; $1\text{ }^{\circ}\text{C s}^{-1}$, red). **b**, Temperature–response profile of the cold-evoked currents shown in **a**. **c**, Response profiles of cold-evoked currents in seven independent oocytes in the absence (black) or presence of a sub-activating concentration of menthol ($20\text{ }\mu\text{M}$; green). **d**, Inward currents evoked in a menthol receptor-expressing oocyte by a saturating cold stimulus ($35\text{ to }5\text{ }^{\circ}\text{C}$, blue trace) were smaller than those evoked by a maximal dose of menthol at room temperature ($500\text{ }\mu\text{M}$, green bar). **e**, Current–voltage relationship for a stimulus evoked by cold ($14\text{ }^{\circ}\text{C}$) in HEK293 cells transfected with the menthol receptor in the absence (black) or presence (green) of a sub-activating dose ($10\text{ }\mu\text{M}$) of menthol. Menthol-induced potentiation and outward rectification of cold-evoked currents are also evident in the accompanying current traces (above) obtained at various voltage steps ($-120\text{ to }70\text{ mV}$). **f**, Current–voltage relationship in transfected HEK293 cells for basal current at $22\text{ }^{\circ}\text{C}$ before (red) and after (blue) warming to $31\text{ }^{\circ}\text{C}$.

Discussion

Menthol has long been known to evoke a sensation of cold. Our findings now provide a molecular explanation for this psychophysical response by demonstrating that cooling compounds and cold are detected by the same molecular entity on primary afferent neurons of the somatosensory system. Moreover, our results show that thermosensation is mediated by a common molecular mechanism that uses TRP ion channels as primary transducers of thermal stimuli. As few as three ion channels (CMR1, VR1 and VRL-1) may provide coverage for a remarkably wide range of temperatures (8–28, >43 and >50 °C, respectively) (Fig. 7a). Still, these channels do not respond to all commonly experienced temperatures, such as ultra-cold (<8 °C) or warm (~30–40 °C), suggesting that additional molecules or mechanisms are involved in mediating thermosensation in these temperature ranges.

If one considers 15 °C as an approximate demarcation between cool and cold^{10,34}, then CMR1 clearly enables detection of temperatures that encompass all of the innocuous cool (15–28 °C) and part of the noxious cold (8–15 °C) range. Furthermore, CMR1 could

contribute to depolarization of fibres at temperatures in the ultra-cold range (<8 °C) if the channel is modulated in a manner that extends its sensitivity range *in vivo*. Indeed, several TRP channels are regulated by receptors that couple to phospholipase C (PLC), and the thermal activation threshold for VR1 can be markedly shifted to lower temperatures by inflammatory agents (for example, bradykinin and nerve growth factor) that activate this pathway^{26,28,35,36}. Other inflammatory products, such as protons and lipids, seem to interact with VR1 directly^{37–39}. Thus, CMR1 might be modulated in a similar manner, expanding its range of temperature sensitivity under normal or pathophysiological conditions. CMR1 may also work in concert with cold-sensitive background potassium channels, which have been suggested to alter the duration or kinetics of cold-evoked action potentials¹⁸. Although CMR1 can clearly underlie activity in the relatively small sub-population of cool/menthol-sensitive fibres, it cannot account for the reportedly large proportion of nociceptive afferent fibres that respond to sub-zero temperatures^{13–15} because its expression is restricted to no more than 15% of trigeminal or DRG neurons. If these latter responses to ultra-cold stimuli are, indeed, physiologically relevant, then they must be mediated by another transducer(s). Finally, it is interesting to note that, unlike CMR1, the heat-sensing receptors VR1 and VRL-1 are activated only in the noxious (pain-producing) range. These biophysical differences in channel sensitivities may help to explain why psychophysical thresholds for cold-evoked pain are not as distinct as they are for heat.

When expressed together, CMR1 and VR1 can endow a cell with distinct thermal thresholds and temperature response ranges for cold and hot, respectively (Fig. 7b). Our calcium-imaging data suggest that a significant proportion (~50%) of CMR1-expressing small-diameter neurons also express VR1 and can therefore be categorized as cold- and heat-responsive nociceptors. These observations provide a plausible molecular explanation for the so-called 'paradoxical' activation of cool (25 °C)-sensitive thermoreceptors by noxious heat^{40,41}, and for the fact that noxious cold is sometimes perceived as burning pain⁴². However, they also raise two interesting questions: how do we distinguish hot from cold, and why doesn't activation of these neurons by an innocuous cool stimulus always cause discomfort or pain? This conundrum might be explained at the cellular level if activation of CMR1 by a cold stimulus elicits action potentials of lower firing frequencies and/or reduced Ca²⁺ influx than VR1-mediated heat responses. Indeed, noxious heat is reported to be a much stronger excitatory stimulus for cat trigeminal C fibres than noxious cold¹¹.

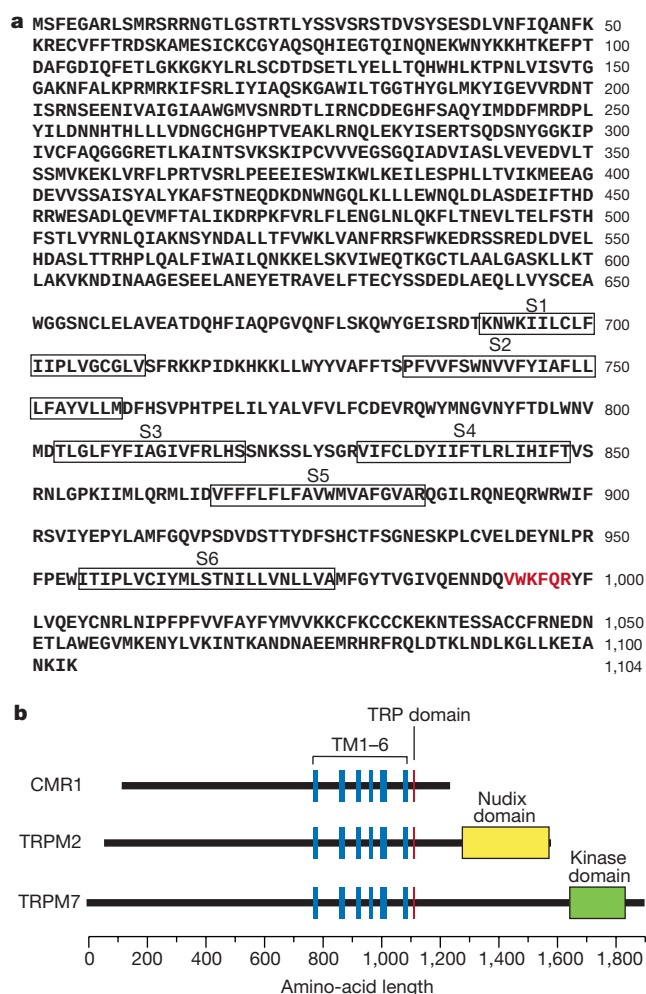


Figure 5 CMR1 is a member of the TRP family of ion channels. **a**, The predicted amino acid sequence determined from the CMR1 cDNA. Boxes designate predicted transmembrane domains, and amino acids encompassing the conserved TRP family motif are shown in red. **b**, Schematic comparison of CMR1 with other TRPM family members, TRPM2 and TRPM7. Proteins are aligned by putative transmembrane domains (blue boxes) and the TRP motif (red boxes) as landmarks. The numeric label is based on the TRPM7 sequence. CMR1 has a significantly shorter C-terminal tail and does not contain any conserved domains indicative of enzymatic activity associated with TRPM2 (ADP ribose pyrophosphatase, Nudix motif, yellow box) or TRPM7 (protein kinase, green box).

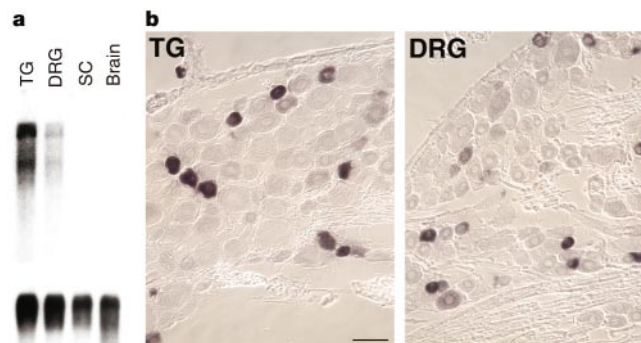


Figure 6 CMR1 is expressed by small-diameter neurons in trigeminal and dorsal root ganglia. **a**, Poly(A)⁺ RNA from adult rat trigeminal ganglia (TG), dorsal root ganglia (DRG), spinal cord (SC) and brain were analysed by northern blotting, revealing two predominant transcripts of roughly 6 and 4.5 kilobases. The blot was re-probed with a rat cyclophilin cDNA (bottom) to control for sample loading. **b**, Histological sections from adult rat trigeminal or dorsal root ganglia showed selective staining (brown) with a digoxigenin-labelled CMR1 probe over neurons with small-diameter (~19 µm) cell bodies. Scale bar, 50 µm.

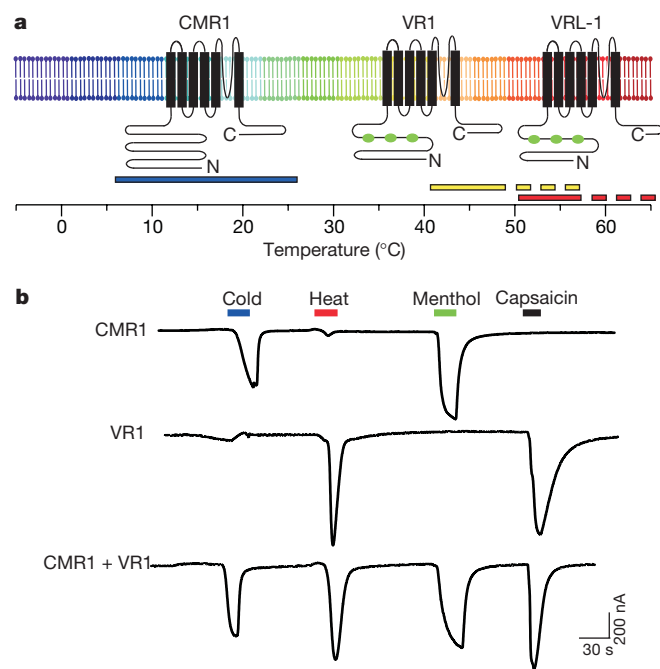


Figure 7 TRP-like channels mediate thermosensation from cold to hot. **a**, Schematic representation of the thermal sensitivity ranges of CMR1, VR1 and VRL-1. The range of each protein's temperature sensitivity is denoted by bars (CMR1, blue; VR1, yellow; VRL-1, red). Other molecules may contribute to temperature sensation in zones not necessarily covered by these channels, in particular warm (30–40°C) or extremely cold (<8°C) zones. **b**, Oocytes co-expressing CMR1 and VR1 demonstrate that these channels are sufficient to confer thermal responsiveness to both cold (menthol) and heat (capsaicin) independently. Bars above traces indicate application of thermal or chemical stimuli (cold, 35 to 8°C; heat, 25 to 50°C; menthol, 100 µM; capsaicin, 1 µM).

Our ability to discriminate among thermal stimuli must also involve decoding of sensory information at the level of the spinal cord and brain, where inputs from afferent fibres having different thermal sensitivities (for example, heat only, cold only, or both) are integrated. For instance, excitatory input from hot- and cold-sensitive (CMR1⁺, VR1⁺) nociceptors may be modulated at the central level by concurrent input from cold-specific (CMR1⁺, VR1[−]) fibres. Such action is supported by two observations. First, central unmasking, or disinhibition of, innocuous cool-induced activity in polymodal C-fibre nociceptors has been proposed as an explanation of the thermal grill illusion⁴², in which simultaneous contact with warm and cool surfaces evokes a sensation of burning pain. Second, an innocuous cool stimulus can evoke the sensation of burning pain in humans after blocking conduction in A fibres or after loss of these fibres from injury (large-fibre neuropathy)^{43,44}. Clearly, answers to these and related questions will require a more detailed understanding of where functionally defined sub-populations of primary afferent neurons make synaptic connections in the spinal cord. The cloning of CMR1 now makes it possible to identify cold-sensitive fibres and to assess the contribution of this ion channel to cold sensation *in vivo*.

When applied to skin or mucous membranes, menthol produces a cooling sensation, inhibits respiratory reflexes, and, at high doses, elicits a pungent or irritant effect that is accompanied by local vasodilation^{25,45}. Most, if not all, of these physiological actions can be explained by excitation of sensory nerve endings, but CMR1 receptors elsewhere might contribute to these or other effects of cooling compounds or cold. For example, intravenous administration of icilin induces intense shivering in mammals⁴⁶, and although this may be mediated by a sympathetic reflex in response to sensory nerve stimulation, it will be interesting to determine whether CMR1 or homologues are expressed in brain regions, such

as the hypothalamus, that regulate core body temperature. Outside of the nervous system, menthol has been reported to increase intracellular calcium in canine tracheal epithelium⁴⁷, an observation that is intriguing in light of the observation that TRPM8 transcripts are expressed in normal prostate epithelium³³. Physiological roles for TRP channels in the prostate are currently unknown, but expression or repression of several TRP genes in tumour cells suggests that these proteins influence cell proliferation, possibly through their ability to regulate intracellular calcium levels²⁶. Because cold is unlikely to be the natural stimulus for TRPM8 in this context, other modulators of the channel may exist, such as an endogenous menthol-like ligand or a bioactive lipid. In any case, to our knowledge CMR1 is now the first member of the long TRP channel family to be associated with a known physiological function or extracellular ligand, making it an important tool for uncovering contributions of these channels to cell growth and sensory perception. □

Methods

Neuronal cell culture and Ca²⁺ microfluorimetry

Trigeminal ganglia were dissected from newborn Sprague Dawley rats and dissociated with 0.125% collagenase P (Boehringer) solution in CMF Hank's solution at 37°C for 20 min (P0 animals) to 30 min (P4 animals), pelleted, and resuspended in 0.05% trypsin at 37°C for 2 min. Ganglia were triturated gently with a fire-polished Pasteur pipette in culture medium (MEM Eagle's/Earle's BSS with 10% horse serum, vitamins, penicillin–streptomycin, L-glutamine and 100 ng ml^{−1} nerve growth factor 75, Invitrogen), then enriched by density gradient centrifugation as described⁴⁸. Cells were resuspended in culture medium and plated onto coverslips coated with polyornithine (PLO) (1 mg ml^{−1}; Sigma) and laminin (5 µg ml^{−1}; Invitrogen). Cultures were examined 1–2 d after plating by Ca²⁺ microfluorimetry as described⁶. For cell selection before patch-clamp analysis, CaCl₂ and pluronic acid were omitted from the loading buffer.

Mammalian cell electrophysiology

Trigeminal neurons responding to 50 µM menthol with an increase in intracellular Ca²⁺ were selected for patch-clamp recordings. HEK293 cells were cultured in DMEM with 10% fetal bovine serum and co-transfected (Lipofectamine 2000, Invitrogen) with 1 µg CMR1 plasmid and 0.1 µg enhanced green fluorescence reporter plasmid to identify transfected cells. Cells were plated onto PLO-coated coverslips the next day and examined 2 d after transfection. Gigaseals were formed with pipettes (Garner Glass 7052, internal diameter 1.1 mm, outer diameter 1.5 mm) having a resistance of 3–5 MΩ in standard pipette solution. Liquid junction potentials (measured in separate experiments) did not exceed 3 mV and thus no correction for this offset was made. Whole-cell voltage clamp was performed at a holding potential of −60 mV with a 200-ms voltage ramp from −120 mV to +80 mV at 3.6 Hz. Data were acquired using Pulse and Pulsefit (HEKA GmbH) software. Recordings were sampled at 20 kHz and filtered at 2 kHz. Pipette and recording solutions for neuronal and mammalian cell experiments are detailed in Supplementary Information. Recordings were performed at 22°C unless noted otherwise. Temperature ramps were generated by cooling or heating the perfusate in a jacketed coil (Harvard) connected to a thermostat. Temperature in the proximity of the patched cell was measured with a miniature thermocouple (MT-29/2, Physitemp) and sampled using an ITC-18 A/D board (Instrutech) and Pulse software. Permeability ratios for monovalent cations to Na (P_X/P_{Na}) were calculated according to $P_X/P_{Na} = \exp(\Delta E_{rev}(Na-X)/RT)$, where $\Delta E_{rev}(Na-X)$ is the reversal potential change, F the Faraday constant, R the universal gas constant and T the absolute temperature. For measurements of Ca²⁺ permeability, P_{Ca}/P_{Na} was calculated according to ref. 49 (equation A6).

Cloning, northern blot and *in situ* hybridization

Trigeminal neurons from newborn rats were dissociated and enriched as described⁴⁸. Polyadenylated RNA (~2 µg) was isolated from these cells (PolyATract Kit, Promega) and used to construct a cDNA library in pcDNA3 (Invitrogen) as described³⁰. Library sub-pools consisting of about 10,000 clones were transiently transfected into HEK293 cells by lipofection and split 24 h later into 8-well glass chamber slides coated with Matrigel (Becton Dickinson). Responses to chemical or thermal stimuli were assessed 6–24 h later using Fura-2 Ca²⁺ imaging. Iterative subdivision and re-screening of a positive pool yielded a single menthol-responsive clone. Northern blotting and *in situ* hybridization histochemistry were performed as described⁶ using the entire CMR1 cDNA to generate ³²P- and digoxigenin-labelled probes, respectively.

Oocyte electrophysiology

Complementary RNA transcripts were synthesized and injected into *Xenopus laevis* oocytes as described⁵⁰. Two-electrode voltage-clamp recordings were performed 2–7 d after injection. Dose-response curves for cooling compounds were performed at room temperature (22–24°C). Icilin (AG-3-5) was provided by E. Wei. Temperature ramps were generated by heating (~35°C) or cooling (~4°C) the perfusate in a Harvard coil and monitoring temperature changes with a thermistor placed near the oocyte.

Received 4 December 2001; accepted 23 January 2002.

Published online 10 February 2002, DOI 10.1038/nature719.

1. Fields, H. L. *Pain* 13–78 (McGraw-Hill, New York, 1987).
2. Julius, D. & Basbaum, A. I. Molecular mechanisms of nociception. *Nature* **413**, 203–210 (2001).
3. Raja, S. N., Meyer, R. A., Ringkamp, M. & Campbell, J. N. in *Textbook of Pain* (ed. Wall, P. D. & Melzack, R.) 11–57 (Churchill Livingstone, Edinburgh, 1999).
4. Nagy, I. & Rang, H. Noxious heat activates all capsaicin-sensitive and also a sub-population of capsaicin-insensitive dorsal root ganglion neurons. *Neuroscience* **88**, 995–997 (1999).
5. Cesare, P. & McNaughton, P. A novel heat-activated current in nociceptive neurons and its sensitization by bradykinin. *Proc. Natl Acad. Sci. USA* **93**, 15435–15439 (1996).
6. Caterina, M. J. *et al.* The capsaicin receptor: a heat-activated ion channel in the pain pathway. *Nature* **389**, 816–824 (1997).
7. Caterina, M. J. *et al.* Impaired nociception and pain sensation in mice lacking the capsaicin receptor. *Science* **288**, 306–313 (2000).
8. Davis, J. B. *et al.* Vanilloid receptor-1 is essential for inflammatory thermal hyperalgesia. *Nature* **405**, 183–187 (2000).
9. Caterina, M. J., Rosen, T. A., Tominaga, M., Brake, A. J. & Julius, D. A capsaicin-receptor homologue with a high threshold for noxious heat. *Nature* **398**, 436–441 (1999).
10. Hensel, H. & Zotterman, Y. The effect of menthol on the thermoreceptors. *Acta Physiol. Scand.* **24**, 27–34 (1951).
11. Bessou, P. & Perl, E. R. Response of cutaneous sensory units with unmyelinated fibers to noxious stimuli. *J. Neurophysiol.* **32**, 1025–1043 (1969).
12. Kress, M., Koltzenburg, M., Reeh, P. W. & Handwerker, H. O. Responsiveness and functional attributes of electrically localized terminals of cutaneous C-fibers in vivo and in vitro. *J. Neurophysiol.* **68**, 581–595 (1992).
13. Simone, D. A. & Kajander, K. C. Excitation of rat cutaneous nociceptors by noxious cold. *Neurosci. Lett.* **213**, 53–56 (1996).
14. Simone, D. A. & Kajander, K. C. Responses of cutaneous A-fiber nociceptors to noxious cold. *J. Neurophysiol.* **77**, 2049–2060 (1997).
15. Cain, D. M., Khasabov, S. G. & Simone, D. A. Response properties of mechanoreceptors and nociceptors in mouse glabrous skin: an in vivo study. *J. Neurophysiol.* **85**, 1561–1574 (2001).
16. Reid, G. & Flonta, M. L. Cold current in thermoreceptive neurons. *Nature* **413**, 480 (2001).
17. Suto, K. & Gotoh, H. Calcium signalling in cold cells studied in cultured dorsal root ganglion neurons. *Neuroscience* **92**, 1131–1135 (1999).
18. Reid, G. & Flonta, M. Cold transduction by inhibition of a background potassium conductance in rat primary sensory neurons. *Neurosci. Lett.* **297**, 171–174 (2001).
19. Askwith, C. C., Benson, C. J., Welsh, M. J. & Snyder, P. M. DEG/ENaC ion channels involved in sensory transduction are modulated by cold temperature. *Proc. Natl Acad. Sci. USA* **98**, 6459–6463 (2001).
20. Pierau, F. K., Torrey, P. & Carpenter, D. O. Mammalian cold receptor afferents: role of an electrogenic sodium pump in sensory transduction. *Brain Res.* **73**, 156–160 (1974).
21. Braun, H. A., Bade, H. & Hensel, H. Static and dynamic discharge patterns of bursting cold fibers related to hypothetical receptor mechanisms. *Pflügers Arch.* **386**, 1–9 (1980).
22. Swandulla, D., Carbone, E., Schafer, K. & Lux, H. D. Effect of menthol on two types of Ca currents in cultured sensory neurons of vertebrates. *Pflügers Arch.* **409**, 52–59 (1987).
23. Schafer, K., Braun, H. A. & Isenberg, C. Effect of menthol on cold receptor activity. Analysis of receptor processes. *J. Gen. Physiol.* **88**, 757–776 (1986).
24. Okazawa, M., Terauchi, T., Shiraki, T., Matsumura, K. & Kobayashi, S. I-Menthol-induced $[Ca^{2+}]_i$ increase and impulses in cultured sensory neurons. *Neuroreport* **11**, 2151–2155 (2000).
25. Eccles, R. Menthol and related cooling compounds. *J. Pharm. Pharmacol.* **46**, 618–630 (1994).
26. Clapham, D. E., Runnels, L. W. & Strubing, C. The TRP ion channel family. *Nature Rev. Neurosci.* **2**, 387–396 (2001).
27. Wei, E. T. & Seid, D. A. AG-3-5: a chemical producing sensations of cold. *J. Pharm. Pharmacol.* **35**, 110–112 (1983).
28. Chuang, H. H. *et al.* Bradykinin and nerve growth factor release the capsaicin receptor from PtdIns(4,5)P₂-mediated inhibition. *Nature* **411**, 957–962 (2001).
29. Kenshalo, D. R. & Duclaux, R. Response characteristics of cutaneous cold receptors in the monkey. *J. Neurophysiol.* **40**, 319–332 (1977).
30. Perraud, A. L. *et al.* ADP-ribose gating of the calcium-permeable LTRPC2 channel revealed by Nudix motif homology. *Nature* **411**, 595–599 (2001).
31. Sano, Y. *et al.* Immunocyte Ca^{2+} influx system mediated by LTRPC2. *Science* **293**, 1327–1330 (2001).
32. Runnels, L. W., Yue, L. & Clapham, D. E. TRP-PLIK, a bifunctional protein with kinase and ion channel activities. *Science* **291**, 1043–1047 (2001).
33. Tsavaler, L., Shaper, M. H., Morkowski, S. & Laus, R. Trp-p8, a novel prostate-specific gene, is up-regulated in prostate cancer and other malignancies and shares high homology with transient receptor potential calcium channel proteins. *Cancer Res.* **61**, 3760–3769 (2001).
34. Rainville, P., Chen, C. C. & Bushnell, M. C. Psychophysical study of noxious and innocuous cold discrimination in monkey. *Exp. Brain Res.* **125**, 28–34 (1999).
35. Harteneck, C., Plant, T. D. & Schultz, G. From worm to man: three subfamilies of TRP channels. *Trends Neurosci.* **23**, 159–166 (2000).
36. Premkumar, L. S. & Ahern, G. P. Induction of vanilloid receptor channel activity by protein kinase C. *Nature* **408**, 985–990 (2000).
37. Zygmunt, P. M. *et al.* Vanilloid receptors on sensory nerves mediate the vasodilator action of anandamide. *Nature* **400**, 452–457 (1999).
38. Hwang, S. W. *et al.* Direct activation of capsaicin receptors by products of lipoxygenases: endogenous capsaicin-like substances. *Proc. Natl Acad. Sci. USA* **97**, 6155–6160 (2000).
39. Jordt, S. E., Tominaga, M. & Julius, D. Acid potentiation of the capsaicin receptor determined by a key extracellular site. *Proc. Natl Acad. Sci. USA* **97**, 8134–8139 (2000).
40. Campero, M., Serra, J., Bostock, H. & Ochoa, J. L. Slowly conducting afferents activated by innocuous low temperature in human skin. *J. Physiol.* **535**, 855–865 (2001).
41. Dodt, E. & Zotterman, Y. The discharge of specific cold fibres at high temperatures (the paradoxical cold). *Acta Physiol. Scand.* **26**, 358–365 (1952).
42. Craig, A. D. & Bushnell, M. C. The thermal grill illusion: unmasking the burn of cold pain. *Science* **265**, 252–255 (1994).
43. Wahren, L. K., Torebjörk, E. & Jorum, E. Central suppression of cold-induced C fibre pain by myelinated fibre input. *Pain* **38**, 313–319 (1989).
44. Yarnitsky, D. & Ochoa, J. L. Release of cold-induced burning pain by block of cold-specific afferent input. *Brain* **113**, 893–902 (1990).
45. Eccles, R. Role of cold receptors and menthol in thirst, the drive to breathe and arousal. *Appetite* **34**, 29–35 (2000).
46. Wei, E. T. Pharmacological aspects of shaking behaviour produced by TRH, AG-3-5, and morphine withdrawal. *Fed. Proc.* **40**, 1491–1496 (1981).
47. Takeuchi, S., Tamaoki, J., Kondo, M. & Konno, K. Effect of menthol on cytosolic Ca^{2+} levels in canine airway epithelium in culture. *Biochem. Biophys. Res. Commun.* **201**, 1333–1338 (1994).
48. Eckert, S. P., Taddese, A. & McCleskey, E. W. Isolation and culture of rat sensory neurons having distinct sensory modalities. *J. Neurosci. Methods* **77**, 183–190 (1997).
49. Lewis, C. A. Ion-concentration dependence of the reversal potential and the single channel conductance of ion channels at the frog neuromuscular junction. *J. Physiol.* **286**, 417–445 (1979).
50. Brake, A. J., Wagenbach, M. J. & Julius, D. New structural motif for ligand-gated ion channels defined by an ionotropic ATP receptor. *Nature* **371**, 519–523 (1994).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com>).

Acknowledgements

We thank the members of our laboratory for encouragement, advice and assistance throughout this project. We are also grateful to R. Nicoll, H. Inghram and A. Basbaum for advice and critical reading of the manuscript. D.D.M. was supported by a National Institutes of Health (NIH) postdoctoral training grant from the University of California—San Francisco (UCSF) Cardiovascular Research Institute and is a recipient of an Arthritis Foundation Postdoctoral Fellowship. W.M.N. was supported by a Fulbright scholarship and a NIH predoctoral training grant from the UCSF Biomedical Sciences Graduate Program. This work was supported by grants from the NIH to D.J.

Correspondence and requests for materials should be addressed to D.J. (e-mail: julius@socrates.ucsf.edu). The GenBank accession number for rat CMR1 is AY072788.

Massive star formation in 100,000 years from turbulent and pressurized molecular clouds

Christopher F. McKee^{*†} & Jonathan C. Tan^{†‡}

^{*} Department of Physics, [†] Department of Astronomy, University of California, Berkeley, California 94720, USA

[‡] Princeton University Observatory, Peyton Hall, Princeton, New Jersey 08544, USA

Massive stars (with mass $m_* > 8$ solar masses $M_\odot$) are fundamental to the evolution of galaxies, because they produce heavy elements, inject energy into the interstellar medium, and possibly regulate the star formation rate. The individual star formation time, t_{sf} , determines the accretion rate of the star; the value of the former quantity is currently uncertain by many orders of magnitude^{1–6}, leading to other astrophysical questions. For example, the variation of t_{sf} with stellar mass dictates whether massive stars can form simultaneously with low-mass stars in clusters. Here we show that t_{sf} is determined by the conditions in the star's natal cloud, and is typically $\sim 10^5$ yr. The corresponding mass accretion rate depends on the pressure within the cloud—which we relate to the gas surface density—and on both the instantaneous and final stellar masses. Characteristic accretion rates are sufficient to overcome radiation pressure from $\sim 100 M_\odot$ protostars, while simultaneously driving intense bipolar gas outflows. The weak dependence of t_{sf} on the final mass of the star allows high- and low-mass star formation to occur nearly simultaneously in clusters.

Massive stars form in dense molecular clumps inside molecular clouds⁷. These regions are highly turbulent and are in approximate virial equilibrium, with comparable values of the gravitational energy and the kinetic energy. Observed star-forming regions have a virial parameter $\alpha_{\text{vir}} = 5\langle\sigma^2\rangle R_{\text{cl}}/GM_{\text{cl}}$ of order unity⁸, where $\langle\sigma^2\rangle$ is the mass-averaged one-dimensional velocity dispersion, R_{cl} the radius, and M_{cl} the mass of the clump. The regions of high-mass star formation studied in ref. 7 are characterized by masses $M_{\text{cl}} \approx 3,800 M_\odot$ and radii $R_{\text{cl}} \approx 0.5$ pc. The corresponding mean column density and visual extinction are $\Sigma \approx 1 \text{ g cm}^{-2}$ and $A_V = (N_{\text{H}}/2 \times 10^{21} \text{ cm}^{-2}) \text{ mag} = 214 \Sigma \text{ mag}$, with a dispersion of a factor of a few. These column densities are far greater than those associated with regions of low-mass star formation, but are comparable to the central surface density in the Orion nebula cluster⁹, which is about 1 g cm^{-2} . These regions have very high mean pressures $\bar{P}$:

$$\frac{\bar{P}}{k_B} = \frac{3M_{\text{cl}}\langle\sigma^2\rangle}{4\pi R_{\text{cl}}^3 k_B} = \left(\frac{3\pi f_{\text{gas}} \alpha_{\text{vir}}}{20k_B}\right) G \Sigma^2 \approx 2.28 \times 10^8 f_{\text{gas}} \alpha_{\text{vir}} \Sigma^2 \text{ K cm}^{-3} \quad (1)$$

where k_B is Boltzmann's constant and f_{gas} is the fraction of the cloud's mass that is in gas, as opposed to stars. They have power-law density profiles¹⁰ $\rho \propto r^{-k_p}$, with $k_p \approx 1.5 \pm 0.5$.

Molecular clouds are observed to be inhomogeneous on a wide range of scales, and numerical simulations show that this is a natural outcome of supersonic turbulence¹¹. Dense, self-gravitating inhomogeneities that are destined to become stars are termed cores. We assume that the distribution of core masses determines the resulting distribution of stellar masses (the initial mass function), and we take this distribution as given. Because of the disruptive effects of protostellar outflows, only a fraction ϵ_{core} of the total core mass M_{core} ends up in the star: $m_{\text{sf}} = \epsilon_{\text{core}} M_{\text{core}}$, where m_{sf} is the final stellar mass (for binary and other multiple star systems, m_{sf} is the total mass of stars in the system). To estimate ϵ_{core} for massive stars,

we assume that their outflows are scaled versions of the outflows from low-mass stars, which is qualitatively consistent with observation¹². For low-mass stars, protostellar outflows typically eject 25–75% of the core mass¹³. A similar analysis applied to high-mass stars yields comparable results (J.C.T. and C.F.M., manuscript in preparation), so we adopt $\epsilon_{\text{core}} = 0.5$ as a typical value.

Our fundamental assumption is that star-forming clumps and the cores embedded within them are each part of a self-similar, self-gravitating turbulent structure that is virialized—that is, in approximate hydrostatic equilibrium—on all scales above the thermal Jeans mass. The density and pressure are then power laws in radius (for the pressure, $P \propto r^{-k_p}$), so that the cloud is a polytrope with $P \propto \rho^{\gamma_p}$. In hydrostatic equilibrium¹⁴, $k_p = 2/(2 - \gamma_p)$ and $k_p = \gamma_p k_\rho = 2\gamma_p/(2 - \gamma_p)$. Let $c \equiv (P/\rho)^{1/2} \propto r^{(1-\gamma_p)/(2-\gamma_p)}$ be the effective sound speed; if the pressure is dominated by turbulent motions, then c is proportional to the line width. Molecular clouds and cloud cores satisfy a line width–size relation in which c increases outwards¹⁵, corresponding to $\gamma_p < 1$. The equation of hydrostatic equilibrium gives $M = k_p c^2 r/G$ for the mass inside r , and $\rho = Ac^2/(2\pi Gr^2)$ for the density at r , where $A = \gamma_p(4 - 3\gamma_p)/(2 - \gamma_p)^2$. It is then possible to determine the properties of a core in terms of the pressure at its surface and the mass of the star that will form within it (see Fig. 1). The radius of a core is then $r_s = 0.074(m_{\text{sf}}/30M_\odot)^{1/2} \Sigma^{-1/2}$ pc; recall that the typical clump observed in ref. 7 has a radius of 0.5 pc, which is set by the angular resolution of those observations. The mean density in a core is $\bar{n}_{\text{H}} = 1.0 \times 10^6 (m_{\text{sf}}/30M_\odot)^{-1/2} \Sigma^{3/2} \text{ cm}^{-3}$, and the r.m.s. velocity dispersion is $1.65(m_{\text{sf}}/30M_\odot)^{1/4} \Sigma^{1/4} \text{ km s}^{-1}$. The thermal sound speed at temperature T is $0.3(T/30\text{K})^{1/2} \text{ km s}^{-1}$, and so a high-mass core is dominated by supersonic turbulence and should be very clumpy.

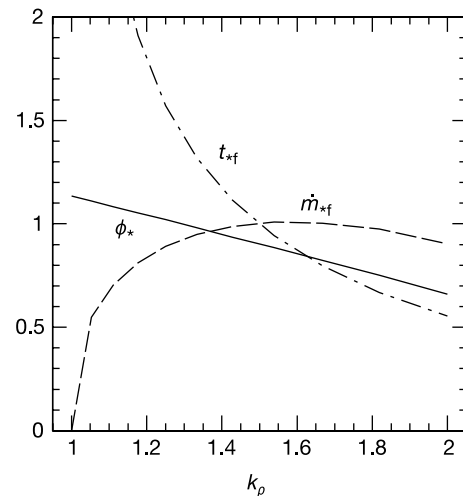


Figure 1 Variation of model parameters and results with k_p . Let $P_s = P_s/\bar{P}$ be the surface pressure of a core. The properties at the surface of the core are given by $r_s = [AGm_{\text{sf}}^2/(2\pi k_p^2 \epsilon_{\text{core}}^2 P_s)]^{1/4}$, $\rho_s = [Ak_p^2 \epsilon_{\text{core}}^2 P_s^3/(2\pi G^3 m_{\text{sf}}^2)]^{1/4}$, and $c_s = [2\pi G^3 m_{\text{sf}}^2 P_s/(Ak_p^2 \epsilon_{\text{core}}^2)]^{1/8}$. We anticipate that the overall star formation efficiency in the clump will be relatively high, so in equation (1) we adopt $f_{\text{gas}} = 2/3$ as a fiducial value. We estimate $\phi_* = P_s/P \approx 2$ (C.F.M. and J.C.T., manuscript in preparation), and we set α_{vir} equal to unity⁷. We take $k_p = 1.5$ as a typical value, corresponding to $\gamma_p = 2/3$, $k_p = 1$ and $A = 3/4$. Following equation (3), we evaluate ϕ_* using the results of ref. 2: $\phi_* = \pi\sqrt{3}[(2 - \gamma_p)^2(4 - 3\gamma_p)]^{(7 - 6\gamma_p)/2} 8^{(3\gamma_p - 5)/2} m_0^{1/(4 - 3\gamma_p)}$, where m_0 is tabulated in ref. 2. Over the entire range of γ_p and k_p relevant to molecular clouds ($0 \leq \gamma_p \leq 1$, $1 \leq k_p \leq 2$), $\phi_* \approx 1.62 - 0.48k_p$, to within about 1%. The star formation time decreases from $3.5t_{\text{ff}}$ to $1.5t_{\text{ff}}$ as γ_p varies from 0 to 1. The variation of t_{sf} and $\dot{m}_{\text{sf}}$ relative to the $k_p = 1.5$ case is also shown. Note that the singular polytropic model in hydrostatic equilibrium breaks down for $k_p = 1$, $\gamma_p = 0$, because then the pressure gradient vanishes ($k_p = 0$).

We now consider the timescale for a star to form in such a core. On dimensional grounds, we expect the protostellar accretion rate to be¹⁶:

$$\dot{m}_* = \phi_* \frac{m_*}{t_{\text{ff}}} \quad (2)$$

so long as radiation pressure is not important. Here m_* is the instantaneous stellar mass, $t_{\text{ff}} = [3\pi/(32G\rho)]^{1/2}$ is the free-fall time and ϕ_* is a dimensionless constant of order unity. This equation could be violated in the sense that $\phi_* \gg 1$ only in the unlikely case that the star forms from a coherent spherical implosion; if the star formation is triggered by an external increase in pressure, ϕ_* could be increased somewhat, but deviations from spherical symmetry in the triggering impulse and in the protostellar core will generally prevent ϕ_* from becoming too large. It could be violated in the opposite sense that $\phi_* \ll 1$ if the core is magnetically dominated, so that collapse could not begin until the magnetic field diffused out of the core. However, magnetic fields are not observed to be dominant in molecular cores¹⁷. Equation (2) implies that the accretion rate, and thus the star formation time $t_{\text{sf}} \propto m_*/\dot{m}_*$, depends weakly on the properties of the ambient medium, $\dot{m}_* \propto \rho^{1/2}$.

We now show that if the collapse is spherical and self-similar, then ϕ_* is quite close to unity provided that radiation pressure does not disrupt the flow. Although we assume the collapse is spherical far from the star, it will naturally proceed via a disk close to the star owing to the angular momentum of the accreting material; we assume this does not limit the flow of matter onto the star as otherwise the disk would become very massive and gravitationally unstable. Shielding by the disk reduces the importance of radiation pressure on the accretion flow, and allows the formation of massive stars provided that the accretion rate is sufficiently large¹⁸. Under the assumption of a polytropic structure in hydrostatic equilibrium, equation (2) implies:

$$\dot{m}_* = \phi_* \epsilon_{\text{core}} \frac{4}{\pi\sqrt{3}} k_p A^{1/2} \frac{c^3}{G} \quad (3)$$

where the value of ρ entering t_{ff} is evaluated at the radius in the initial cloud that just encloses the gas that goes into the star when its mass is m_* . For the isothermal case ($\gamma_p = 1$) and for $\epsilon_{\text{core}} = 1$, this reduces to the classic result of ref. 19 with $\phi_* = 0.975\pi\sqrt{3}/8 = 0.663$. The accretion rate $\dot{m}_*$ is² proportional to $t^{3-3\gamma_p}$, so that for $\gamma_p < 1$, as observed, the accretion rate accelerates^{2,4}. As discussed in ref. 2, termination of the accretion breaks the self-similarity once the expansion wave encloses m_{sf} , which occurs at $t \approx 0.45t_{\text{sf}}$. Thereafter, the relation $\dot{m}_* \propto t^{3-3\gamma_p}$ becomes approximate, but the approximation should be reasonably good². The star formation time is then given by:

$$t_{\text{sf}} = \frac{(4 - 3\gamma_p)}{\phi_*} t_{\text{ff}} \quad (4)$$

Evaluating ϕ_* (see Fig. 1), we conclude that $\dot{m}_* \approx m_*/t_{\text{ff}}$ to within a factor 1.5 for spherical cores in which the effective sound speed increases outward.

We can express the accretion rate in terms of the mean pressure of the clump in which the star forms, $\bar{P}$ (equation (1)), and the final mass of the star, m_{sf} :

$$\dot{m}_* = 4.75 \times 10^{-4} \epsilon_{\text{core}}^{1/4} (f_{\text{gas}} \phi_p \alpha_{\text{vir}})^{3/8} \left(\frac{m_{\text{sf}}}{30M_{\odot}} \right)^{3/4} \Sigma^{3/4} \left(\frac{m_*}{m_{\text{sf}}} \right)^j M_{\odot} \text{ yr}^{-1} \quad (5)$$

where $\phi_p \equiv P_s/\bar{P}$ is the ratio of the core's surface pressure to the mean pressure in the clump, and $j \equiv 3(2 - k_p)/[2(3 - k_p)]$; for $k_p = 3/2$, $j = 1/2$. This typical accretion rate is large because the core is embedded in a high-pressure environment, so that in hydrostatic equilibrium it has a high density and short free-fall time. Because massive cores are turbulent and clumpy, we expect the accretion rate to exhibit large fluctuations. Whereas clumpiness is

an attribute of massive star-forming regions in our model, it is not a prerequisite, as suggested in ref. 3. The star formation time is:

$$t_{\text{sf}} = 1.26 \times 10^5 \epsilon_{\text{core}}^{-1/4} (f_{\text{gas}} \phi_p \alpha_{\text{vir}})^{-3/8} \left(\frac{m_{\text{sf}}}{30M_{\odot}} \right)^{1/4} \Sigma^{-3/4} \text{ yr} \quad (6)$$

The weak dependence of the star formation time on the final stellar mass means that low-mass and high-mass stars can form approximately at the same time. Furthermore, t_{sf} is small compared to the estimated timescale ($\sim 10^6$ yr) for cluster formation²⁰. With the fiducial values of the parameters, a $100M_{\odot}$ star forming in a clump with $\Sigma = 1 \text{ g cm}^{-2}$ has a final accretion rate of $1.1 \times 10^{-3} M_{\odot} \text{ yr}^{-1}$ and a star formation time of 1.8×10^5 yr.

A necessary condition for massive star formation is that the ram pressure associated with accretion exceed the radiation pressure at the point where the dust in the infalling gas is destroyed²¹. For a $100M_{\odot}$ star, this requires¹⁸ $\dot{m}_* \geq 6 \times 10^{-4} M_{\odot} \text{ yr}^{-1}$, which is satisfied for $\Sigma \geq 0.5 \text{ g cm}^{-2}$. Once the core has formed a star, the star can continue to grow by Bondi-Hoyle accretion²² provided $m_* \lesssim 10M_{\odot}$ so that radiation pressure does not prevent the focusing of gas streamlines in the wake of the star. The Bondi-Hoyle accretion rate $\dot{m}_* = 2.2 \times 10^{-7} (M_{\odot}/4 \times 10^3 M_{\odot})^{-5/4} \Sigma^{3/4} (m_*/10M_{\odot})^2 M_{\odot} \text{ yr}^{-1}$ (C.F.M. and J.C.T., manuscript in preparation) is so low that this process does not significantly alter the stellar mass under the conditions observed in the clumps of ref. 7.

Direct comparison of our results with observation is difficult

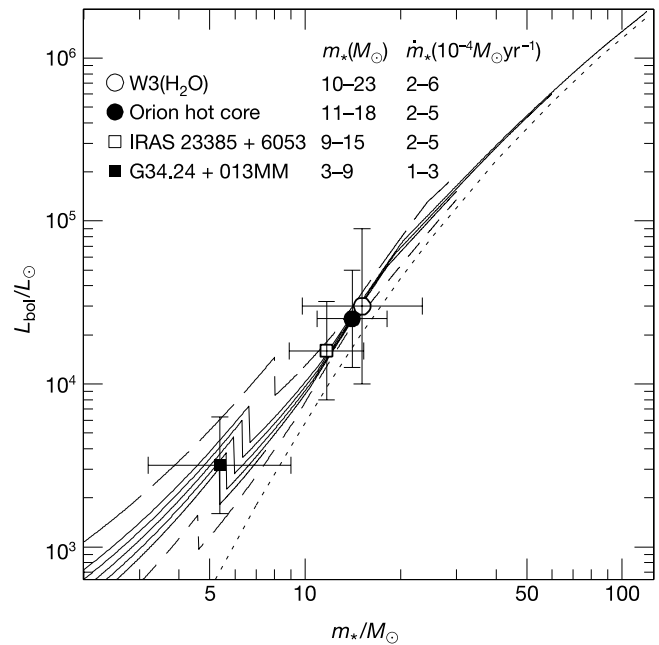


Figure 2 Derived properties of nearby massive protostars. Solid lines show the predicted evolution in luminosity of protostars (including their accretion disks, but allowing for the powering of protostellar outflows so that $f_{\text{acc}} = 0.5$; J.C.T. and C.F.M., manuscript in preparation) of final mass 7.5, 15, 30, 60 and $120M_{\odot}$ accreting from cores with $k_p = 1.5$ embedded in a $\Sigma = 1 \text{ g cm}^{-2}$ clump, typical of Galactic regions⁷. The luminosity step occurring at around $5\text{--}7M_{\odot}$, depending on the model, corresponds to the onset of deuterium shell burning, which swells the protostellar radius by a factor of about two and thus reduces the accretion luminosity by the same factor. The dashed and long-dashed lines show a $30M_{\odot}$ star forming in a clump with mean pressure ten times smaller and larger than the fiducial value, respectively. The dotted line shows the zero-age main-sequence luminosity from ref. 26. Four observed hot molecular cores are shown: G34.24+0.13MM²⁷, IRAS23385+6053²⁸, Orion hot core²⁹ and W3(H₂O)—the Turner–Welch object³⁰. The vertical error bar illustrates the uncertainty in the bolometric luminosity. The horizontal error bar shows the corresponding range of allowed values of m_* for the $\Sigma = 1 \text{ g cm}^{-2}$ models. These values and the constraints on $\dot{m}_*$ are listed here for each source.

because the actual masses of massive protostars are poorly determined. Our approach is to predict the properties of some well studied massive protostars in terms of their bolometric luminosities. The bolometric luminosity L_{bol} has contributions from main-sequence nuclear burning L_{ms} , deuterium burning L_{D} , and accretion L_{acc} . The accretion luminosity $L_{\text{acc}} = f_{\text{acc}} G m_* \dot{m}_* / r_*$, where f_{acc} is a factor of order unity accounting for energy radiated by an accretion disk, advected into the star or converted into kinetic energy of outflows, and where the stellar radius r_* may depend sensitively on the accretion rate $\dot{m}_*$. Massive stars join the main sequence during their accretion phase at a mass that also depends on the accretion rate²³. To treat accelerating accretion rates, we have developed a simple model for protostellar evolution based on that of refs 6 and 24. The model accounts for the total energy of the protostar as it accretes and dissociates matter and, if the central temperature $T_c \geq 10^6$ K, burns deuterium. We have modified this model to include additional processes, such as deuterium shell burning, and we have calibrated these modifications against the more detailed calculations of refs 23 and 25.

Our model allows us to make predictions for the masses and accretion rates of embedded protostars that are thought to power hot molecular cores (C.F.M. and J.C.T., manuscript in preparation). Figure 2 compares our theoretical tracks with the observed bolometric luminosities of several sources. We find that uncertainties in the value of the pressure create only small uncertainties in m_* for L_{bol} in excess of a few times 10^4 solar luminosities.

The infrared and submillimetre spectra of accreting protostars and their surrounding envelopes have been modelled in ref. 5, modelling the same sources shown in Fig. 2. We note that uncertainties in the structure of the gas envelope and the possible contributions from additional surrounding gas cores or diffuse gas will affect the observed spectrum. Comparing results, our inferred stellar masses are similar, but our accretion rates are systematically smaller by factors of ~ 2 – 5 . The modelled⁵ high accretion rates of $\sim 10^{-3} M_{\odot} \text{yr}^{-1}$ for stars with $m_* \approx 10 M_{\odot}$ would be difficult to achieve unless the pressure was increased substantially; for example, if the stars are destined to reach $m_{\text{sf}} \approx 30 M_{\odot}$, pressure increases of a factor ~ 40 are required. $\square$

Received 17 September 2001; accepted 2 January 2002.

- Bernasconi, P. A. & Maeder, A. About the absence of a proper zero age main sequence for massive stars. *Astron. Astrophys.* **307**, 829–839 (1996).
- McLaughlin, D. E. & Pudritz, R. E. Gravitational collapse and star formation in logotropic and nonisothermal spheres. *Astrophys. J.* **476**, 750–765 (1997).
- Stahler, S. W., Palla, F. & Ho, P. T. P. in *Protostars & Planets IV* (eds Mannings, V., Boss, A. P. & Russell, S. S.) 327–351 (Univ. Arizona Press, Tucson, 2000).
- Behrend, R. & Maeder, A. Formation of massive stars by growing accretion rate. *Astron. Astrophys.* **373**, 190–198 (2001).
- Osorio, M., Lizano, S. & D'Alessio, P. Hot molecular cores and the formation of massive stars. *Astrophys. J.* **525**, 808–820 (1999).
- Nakano, T., Hasegawa, T., Morino, J.-I. & Yamashita, T. Evolution of protostars accreting mass at very high rates: Is Orion IRC2 a huge protostar? *Astrophys. J.* **534**, 976–983 (2000).
- Plume, R., Jaffe, D. T., Evans, N. J. II, Martin-Pintado, J. & Gomez-Gonzalez, J. Dense gas and star formation: Characteristics of cloud cores associated with water masers. *Astrophys. J.* **476**, 730–749 (1997).
- Bertoldi, F. & McKee, C. F. Pressure-confined clumps in magnetized molecular clouds. *Astrophys. J.* **395**, 140–157 (1992).
- Hillenbrand, L. A. & Hartmann, L. W. A preliminary study of the Orion nebula cluster structure and dynamics. *Astrophys. J.* **492**, 540–553 (1998).
- van der Tak, F. F. S., van Dishoeck, E. F., Evans, N. J. II & Blake, G. A. Structure and evolution of the envelopes of deeply embedded massive young stars. *Astrophys. J.* **537**, 283–303 (2000).
- Vázquez-Semadeni, E., Ostriker, E. C., Passot, T., Gammie, C. F. & Stone, J. M. in *Protostars & Planets IV* (eds Mannings, V., Boss, A. P. & Russell, S. S.) 3–28 (Univ. Arizona Press, Tucson, 2000).
- Richer, J. S., Shepherd, D. S., Cabrit, S., Bachiller, R. & Churchwell, E. in *Protostars & Planets IV* (eds Mannings, V., Boss, A. P. & Russell, S. S.) 867–894 (Univ. Arizona Press, Tucson, 2000).
- Matzner, C. D. & McKee, C. F. Efficiencies of low-mass star and star cluster formation. *Astrophys. J.* **545**, 364–378 (2000).
- McLaughlin, D. E. & Pudritz, R. E. A model for the internal structure of molecular cloud cores. *Astrophys. J.* **469**, 194–208 (1996).
- Larson, R. B. Turbulence and star formation in molecular clouds. *Mon. Not. R. Astron. Soc.* **194**, 809–826 (1981).
- Stahler, S. W., Shu, F. H. & Taam, R. E. The evolution of protostars. I—Global formulation and results. *Astrophys. J.* **241**, 637–654 (1980).
- Crutcher, R. M. Magnetic fields in molecular clouds: Observations confront theory. *Astrophys. J.* **520**, 706–713 (1999).

- Iijima, J. & Adams, F. C. Infall collapse solutions in the inner limit: Radiation pressure and its effects on star formation. *Astrophys. J.* **462**, 874–887 (1996).
- Shu, F. H. Self-similar collapse of isothermal spheres and star formation. *Astrophys. J.* **214**, 488–497 (1977).
- Palla, F. & Stahler, S. W. Star formation in the Orion nebula cluster. *Astrophys. J.* **525**, 772–783 (1999).
- Wolfire, M. G. & Cassinelli, J. Conditions for the formation of massive stars. *Astrophys. J.* **319**, 850–867 (1987).
- Bonnell, I. A., Bate, M. R., Clarke, C. J. & Pringle, J. E. Competitive accretion in embedded stellar clusters. *Mon. Not. R. Astron. Soc.* **323**, 785–794 (2001).
- Palla, F. & Stahler, S. W. The evolution of intermediate-mass protostars. II—Influence of the accretion flow. *Astrophys. J.* **392**, 667–677 (1992).
- Nakano, T., Hasegawa, T. & Norman, C. The mass of a star formed in a cloud core: Theory and its application to the Orion A cloud. *Astrophys. J.* **450**, 183–195 (1995).
- Palla, F. & Stahler, S. W. The evolution of intermediate-mass protostars. I—Basic results. *Astrophys. J.* **375**, 288–299 (1991).
- Schaller, G., Schaerer, D., Meynet, G. & Maeder, A. New grids of stellar models from 0.8 to 120 solar masses at $Z = 0.020$ and $Z = 0.001$. *Astron. Astrophys. Suppl.* **96**, 269–331 (1992).
- Hunter, T. R. et al. G34.24+0.13MM: A deeply embedded proto-B-star. *Astrophys. J.* **493**, L97–L100 (1998).
- Molinari, S., Testi, L., Brand, J., Cesaroni, R. & Palla, F. IRAS 23385+6053: A prototype massive class 0 object. *Astrophys. J.* **505**, L39–L42 (1998).
- Kaufman, M. J., Hollenbach, D. J. & Tielens, A. G. G. M. High-temperature molecular cores near massive stars and application to the Orion hot core. *Astrophys. J.* **497**, 276–287 (1998).
- Wyrowski, F., Schilke, P., Walmsley, C. M. & Menten, K. M. Hot gas and dust in a protostellar cluster near W3(OH). *Astrophys. J.* **514**, L43–L46 (1999).

Acknowledgements

We thank S. Stahler, R. Pudritz, M. Walmsley and M. Krumholz for discussions. This work was supported by the NSF, by NASA (which supports the Center for Star Formation Studies) and (for J.C.T.) by a Spitzer-Cotsen fellowship from Princeton University.

Correspondence and requests for materials should be addressed to C.F.M. (e-mail: cmckee@mckee.berkeley.edu).

Coherent emission of light by thermal sources

Jean-Jacques Greffet*†, Rémi Carminati*, Karl Joulain*, Jean-Philippe Mulet*, Stéphane Mainguy† & Yong Chen‡

* Laboratoire EM2C, CNRS, Grande Voie des Vignes, Châtenay-Malabry 92295 Cedex, France

† CEA CESTA, Le Barp, 33114, France

‡ Laboratoire de Microstructures et de Microélectronique, CNRS, av. H. Ravera, 92220 Bagneux, France

¶ Present address: The Institute of Optics, University of Rochester, New York 14627, USA

A thermal light-emitting source, such as a black body or the incandescent filament of a light bulb, is often presented as a typical example of an incoherent source and is in marked contrast to a laser. Whereas a laser is highly monochromatic and very directional, a thermal source has a broad spectrum and is usually quasi-isotropic. However, as is the case with many systems, different behaviour can be expected on a microscopic scale. It has been shown recently^{1,2} that the field emitted by a thermal source made of a polar material is enhanced by more than four orders of magnitude and is partially coherent at a distance of the order of 10 to 100 nm. Here we demonstrate that by introducing a periodic microstructure into such a polar material (SiC) a thermal infrared source can be fabricated that is coherent over large distances (many wavelengths) and radiates in well defined directions. Narrow angular emission lobes similar to antenna lobes are observed and the emission spectra of the source depends on the observation angle—the so-called Wolf effect^{3,4}. The origin of the coherent emission lies in the diffraction of surface-phonon polaritons by the grating.

It is usually taken for granted that light spontaneously emitted by

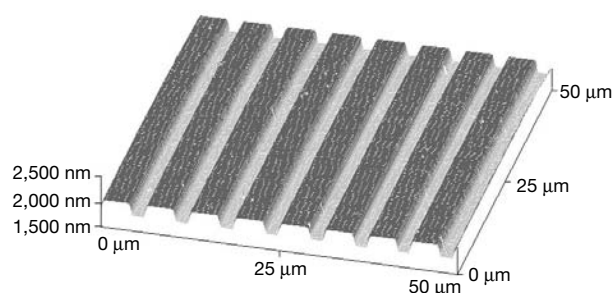


Figure 1 Image of the grating obtained by atomic force microscopy. Its period $d = 0.55\lambda$ ($\lambda = 11.36 \mu\text{m}$) was chosen so that a surface wave propagating along the interface could be coupled to a propagating wave in the range of frequencies of interest. The depth $h = \lambda/40$ was optimized so that the peak emissivity is 1 at $\lambda = 11.36 \mu\text{m}$. It was fabricated on SiC by standard optical lithography and reactive ion-etching techniques.

different points of a thermal source cannot interfere. In contrast, different points of an antenna emit waves that interfere constructively in particular directions producing well defined angular lobes. The intensity emitted by a thermal source is the sum of the intensities emitted by different points so that it cannot be directional. However, it has been shown^{1,2} recently that some planar sources may have a spectral coherence length in the plane much larger than a wavelength and can be quasi-monochromatic in the near-field. This paves the way for the construction of a thermal source that could radiate light within narrow angular lobes as an antenna instead of having the usual quasi-lambertian angular behaviour.

Here we report experimental measurements demonstrating that it is possible to build an infrared antenna by properly designing a periodic microstructure on a polar material. Such an antenna radiates infrared light in a narrow solid angle when it is heated. Another unusual property of this source is that its emission spectrum depends on the observation direction. This property was first predicted by Wolf as a consequence of spatial correlations

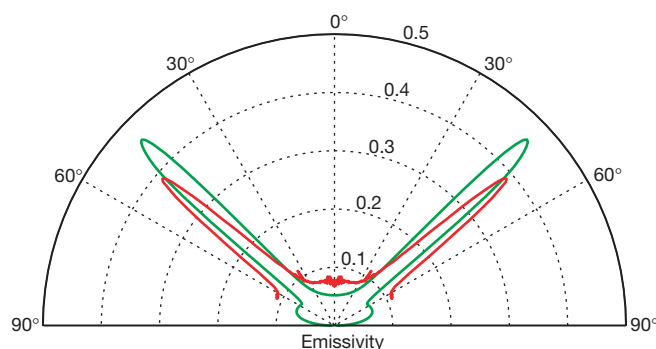


Figure 2 Polar plot of the emissivity of the grating depicted in Fig. 1 at $\lambda = 11.36 \mu\text{m}$ and for p -polarization. Red, experimental data; green, theoretical calculation. The measurements were taken by detecting the intensity emitted by the sample in the far field as a function of the emission angle. A HgCdTe detector placed at the focal length of a ZnSe lens was used. The sample was mounted on a rotation stage. The theoretical result was obtained by computing the reflectivity of the sample and using Kirchhoff's law ($\epsilon = \alpha = 1 - R$). To fit the data, we took into account the spectral resolution ($0.22 \mu\text{m}$) and the angular resolution (3°) of the measurements. The disagreement is due to the fact that for the calculation, the index at room temperature is used whereas emission data were taken with a sample in a local thermal equilibrium situation at a temperature of 773 K. Comparison between the two curves illustrates the validity of Kirchhoff's law for polarized monochromatic directional quantities. The surrounding medium was at 300 K and the background signal was subtracted. The emissivity for s -polarization (not shown) does not show any peak and is very close to its value for a flat surface. Note that most of the emitted light is emitted in the narrow lobe (that is, coherently).

for random sources^{3,4}. This effect has been demonstrated experimentally for artificial secondary sources^{4,5} but has never been observed for direct thermal emission. In addition, the emissivity of the source is enhanced by a factor of 20 compared to the emissivity of a flat surface.

Using theories developed recently⁶ to interpret the emission data by gratings, we have designed and optimized a periodic surface profile that produces a strong peak of the emissivity around a wavelength $\lambda = 11.36 \mu\text{m}$. The grating (see Fig. 1) has been ruled on a SiC substrate. A similar grating of doped silicon with very deep depths has been investigated⁷. The authors attributed the particular properties observed to organ pipe modes in the microstructure⁷. However, the role of coherence induced by surface waves and the exact mechanism were not understood at that time^{2,6}.

The measurements of the thermal emission in a plane perpendicular to the lines of the grating are shown in Fig. 2. Emission is highly directional and looks very similar to the angular pattern of an antenna. We have also plotted (see Fig. 2) the calculated emission pattern. The qualitative discussion of the introduction suggests that the small angular width of the emission pattern is a signature of the local spatial coherence of the source. A proof of this stems from the fact that the source has a width $L = 5 \text{ mm}$ and a spectral coherence length $l \ll L$ and that its temperature is uniform. Hence, we can assume that the source is a quasi-homogeneous source⁸. With this assumption, it is known that the radiant intensity and the spectral degree of coherence in the plane of the source are related by a Fourier transform relationship⁸. Therefore, the angular width θ of the lobe emission varies qualitatively with λ/l for this locally coherent source instead of with λ/L , as for a globally coherent source. Thus a small angular aperture of the far-field radiation is the signature of a spectral coherence length in the source much larger than the wavelength. To overcome the experimental resolution limit of our direct emissivity measurement, we measured the reflectivity R . From Kirchhoff's law⁹, we know that the polarized directional spectral emissivity ϵ is given by $\epsilon = \alpha = 1 - R$ where α is the absorptivity and R is the reflectivity of the grating. Results are plotted in Fig. 3. There is a remarkable quantitative agreement between the data taken at room temperature and theoretical calculations. We note that the peak at 60° has an angular width θ as narrow as 1° so that the corresponding spectral coherence length is as large as $\lambda/\theta \approx 60\lambda \approx 0.6 \text{ mm}$. This suggests that a thermal source with a size L on the order of the spectral coherence length l , namely a globally coherent thermal source, could be achieved.

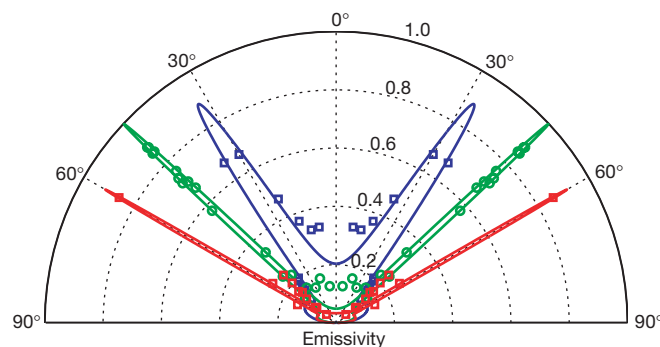


Figure 3 Emissivity of a SiC grating in p -polarization. Blue, $\lambda = 11.04 \mu\text{m}$; red, $\lambda = 11.36 \mu\text{m}$; green, $\lambda = 11.86 \mu\text{m}$. The emissivity was deduced from measurements of the specular reflectivity R using Kirchhoff's law. The data have been taken at ambient temperature using a Fourier transform infrared (FTIR) spectrometer as a source and a detector mounted on a rotating arm. The angular acceptance of the spectrometer was reduced to a value lower than the angular width of the dip. The experimental data are indicated by circles; the lines show the theoretical results. An excellent agreement is obtained when the data are taken at ambient temperature, which supports our interpretation of the slight disagreement in Fig. 1 being due to the variation of index with temperature.

A surprising property of the emissivity patterns of Fig. 3 is that they depend strongly on the wavelength. This suggests that the emission spectra depend on the emission direction. This would be a manifestation of the Wolf effect^{3,4}. To observe this effect, we have taken several spectra of the reflectivity of the surface at different angles. Fig. 4 shows experimental and theoretical spectra for different observation angles. The position of the peaks of emissivity (dips of reflectivity) depends strongly on the observation angle. It is important to emphasize that this property is not merely a scattering effect but is a consequence of the partial spatial coherence of the source. The value of the reflectivity is also remarkable. By ruling a grating onto a material which is essentially a mirror, we were able to produce a perfect absorber. This behaviour has already been observed for metallic gratings and attributed to the resonant excitation of surface plasmons. This is the first time, to our knowledge, that total absorption in the infrared owing to excitation of surface-phonon polaritons has been reported.

In order to prove experimentally the role of the surface wave, we have done spectral measurements of the emissivity for *s*- and *p*-polarization. The peaks are never observed for *s*-polarization nor for *p*-polarization in the spectral region where surface waves cannot exist. In order to characterize quantitatively the role of the surface waves, we have obtained the dispersion relation from reflectivity measurements⁶. The results are displayed in Fig. 5 and compared with theory. We note that the interaction of the surface wave with the grating produces the aperture of a gap close to the band edge. Figure 5 shows that our experiment allows us to directly see surface-phonon polaritons. It also yields additional insight into the Wolf-effect^{3,4} mechanism. Emission of infrared light has already been used to study surface excitations, but using prisms to couple the surface waves to propagating modes¹⁰.

We now discuss the physical origin of coherent thermal emission. We wish to understand how random thermal motion can generate a coherent current along the interface. The key lies in the coherent properties of surface waves (either surface-plasmon polaritons or surface-phonon polaritons) demonstrated in refs 1 and 2. Both are mechanical delocalized collective excitations involving charges. Surface-phonon polaritons are phonons in a polar material, whereas surface-plasmon polaritons are acoustical-type waves in an electron gas. In both cases, these waves have the following two

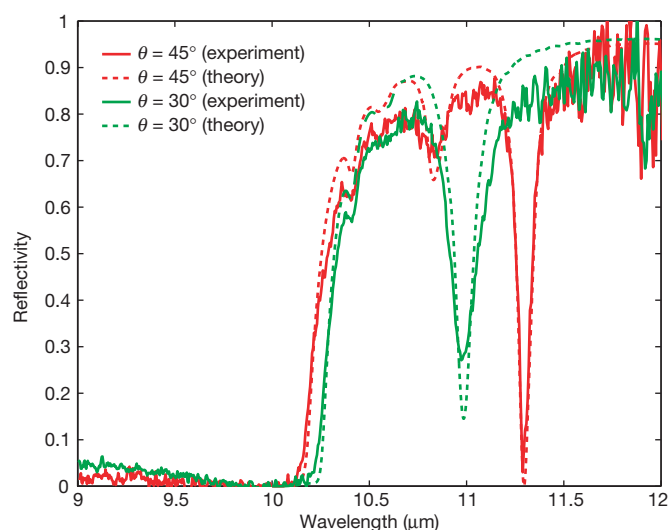


Figure 4 Comparison between measured and calculated spectral reflectivities of a SiC grating at room temperature. The incident light is *p*-polarized. The dip observed at 45° and $\lambda = 11.36 \mu\text{m}$ coincides with the emission peak observed in Figs 2 and 3. The figure shows clearly that the reflectivity spectra depend on the observation angle. Using Kirchhoff's law, it follows that the emission spectra depend on the observation angle. This is a manifestation of the Wolf effect^{3,4}.

properties: (1) they are mechanical modes of the system that can be resonantly excited; (2) they are charge-density waves, that is, they generate electromagnetic fields. Because these excitations are delocalized, so are the corresponding electromagnetic fields.

From a classical point of view, each volume element of the thermal source can be modelled by a random point-like source that excites an extended mode: the surface wave. This is similar to some extent to the emission of sound by a string of a piano. The source is a hammer that strikes the string at a particular point. Then the modes of the string are excited producing a vibration along the full length of the string. At that point of the analogy, as anyone can hear the vibrations of a piano string, we may wonder why the coherent electromagnetic surface waves are not usually observed. The reason is that surface modes have a wavevector larger than $2\pi/\lambda$ so that they are evanescent. Their effect is not seen in the far-field.

However, by ruling a grating on the interface, we are able to couple these surface modes to propagating modes. The relationship between the emission angle θ and the wavelength λ is simply given by the usual grating law

$$\frac{2\pi}{\lambda} \sin\theta = k_{\parallel} + p \frac{2\pi}{d}$$

where p is an integer and $k_{\parallel}$ is the wavevector of the surface wave. Thus, by modifying the characteristics of the surface profile, it is possible to modify the direction and the value of the emissivity of the surface at a given wavelength. It is also possible to modify the emission spectrum in a given direction. Such gratings can be used to design infrared sources with specific properties.

This may also have interesting applications such as modifying the radiative heat transfer for a given material. Indeed, we have demonstrated that a reflectivity of 94% can be reduced to almost zero in the infrared for SiC. This could also be done for glass, which

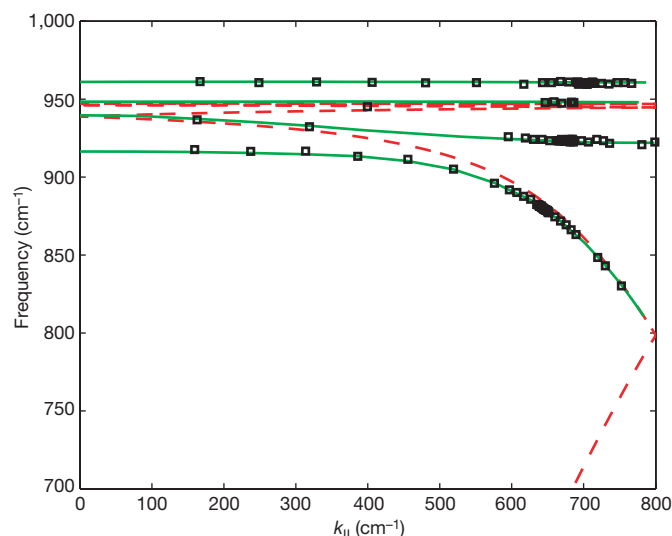


Figure 5 Dispersion relation wavevector, $\omega(k_{\parallel})$, of surface-phonon polaritons. Data points, experimental dispersion relation. Solid green curve, theoretical dispersion relation for the grating. Dotted red curve, theoretical dispersion relation for the flat surface. This figure explains the mechanism of the Wolf effect^{3,4} for this particular source. The spatial coherence in the plane of the source is due to the presence of a surface wave. For a fixed frequency ω , it can be seen that there is only one possible wavevector $k_{\parallel}(\omega)$. Thus the spectral degree of coherence at ω oscillates² with a particular wavelength $2\pi/k_{\parallel}(\omega)$. When observing in the far field at an angle θ such that $ck_{\parallel}(\omega)/\omega = \sin\theta$ there is a strong contribution of the surface wave at frequency ω . By varying the observation angle, the frequency varies according to the dispersion relation of the surface wave. It is seen that the strong Wolf effect produced by this source is due to (1) the thermal excitation of surface waves which produce the spatial coherence and (2) the propagation in vacuum which selects one particular wavevector.

is a polar material that has a large reflectivity in the infrared owing to the presence of resonances. This would allow us to increase the radiative cooling of the material if the emission is enhanced in a region where absorption is low, because the atmosphere does not emit. Another promising application of our results is the possibility of modifying the heat transfer in the near-field. Materials that are separated by distances smaller than the typical wavelength exchange radiative energy through evanescent waves. When surface waves are resonantly excited, they provide the leading contribution¹¹. Thus, the heat transfer is almost monochromatic. This may be used to enhance the efficiency of infrared photovoltaic cells¹². □

Received 25 June; accepted 20 December 2001.

- Shchegrov, A., Joulain, K., Carminati, R. & Greffet, J. J. Near-field spectral effects due to electromagnetic surface excitations. *Phys. Rev. Lett.* **85**, 1548–1551 (2000).
- Carminati, R. & Greffet, J. J. Near-field effects in spatial coherence of thermal sources. *Phys. Rev. Lett.* **82**, 1660–1663 (1999).
- Wolf, E. Non-cosmological red-shifts of spectral lines. *Nature* **326**, 363–365 (1987).
- Wolf, E. & James, D. F. Correlation-induced spectral changes. *Rep. Prog. Phys.* **59**, 771–818 (1996).
- Morris, G. M. & Faklis, D. Effects of source correlation on the spectrum of light. *Opt. Commun.* **62**, 5–11 (1987).
- Le Gall, J., Olivier, M. & Greffet, J. J. Experimental and theoretical study of reflection and coherent thermal emission by a SiC grating supporting a surface photon polariton. *Phys. Rev. B* **55**, 10105–10114 (1997).
- Hesketh, P. J., Zemel, J. N. & Gebhart, B. Organ pipe radiant modes of periodic micromachined silicon surfaces. *Nature* **325**, 549–551 (1986).
- Mandel, L. & Wolf, E. *Optical Coherence and Quantum Optics*. Sec. 5.3 (Cambridge Univ. Press, New York, 1995).
- Greffet, J. J. & Nieto-Vesperinas, M. Field theory for the generalized bidirectional reflectivity: derivation of Helmholtz's reciprocity principle and Kirchhoff's law. *J. Opt. Soc. Am. A* **10**, 2735–2744 (1998).
- Zhizhin, G. N., Vinogradov, E. A., Moskalova, M. A. & Yakovlev, V. A. Applications of surface polaritons for vibrational spectroscopic studies of thin and very thin films. *Appl. Spectrosc. Rev.* **18**, 171–263 (1982).
- Mulet, J. P., Joulain, K., Carminati, R. & Greffet, J. J. Nanoscale radiative heat transfer between a small particle and a plane surface. *Appl. Phys. Lett.* **78**, 2931–2933 (2001).
- Whale, M. D. in *Proc. Conf. on 'Heat Transfer and Transport Phenomena in Microscale'* (ed. Celata, G. P.) 339–346 (Begell House, New York, 2000).

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.-J.G.
(e-mail: greffet@em2c.epc.fr).

A general process for *in situ* formation of functional surface layers on ceramics

Toshihiro Ishikawa, Hiroyuki Yamaoka, Yoshikatsu Harada, Teruaki Fujii & Toshio Nagasawa

Ube Research Laboratory, Corporate Research & Development,
Ube Industries Ltd, 1978-5 Kogushi, Ube City, Yamaguchi 755-8633, Japan

Ceramics are often prepared with surface layers of different composition from the bulk^{1,2}, in order to impart a specific functionality to the surface or to act as a protective layer for the bulk material^{3,4}. Here we describe a general process by which functional surface layers with a nanometre-scale compositional gradient can be readily formed during the production of bulk ceramic components. The basis of our approach is to incorporate selected low-molecular-mass additives into either the precursor polymer from which the ceramic forms, or the binder polymer used to prepare bulk components from ceramic powders. Thermal treatment of the resulting bodies leads to controlled phase separation ('bleed out') of the additives, analogous to the normally

undesirable outward loss of low-molecular-mass components from some plastics^{5–9}; subsequent calcination stabilizes the compositionally changed surface region, generating a functional surface layer. This approach is applicable to a wide range of materials and morphologies, and should find use in catalysts, composites and environmental barrier coatings.

To avoid the concentration of thermomechanical stress at the interface between the surface layer and the bulk material, many materials have been developed that have gradually varying properties as the distance into the material increases¹⁰. Such materials can contain gradients in morphology or in composition. Gradients in morphology can, for example, result in materials that have a graded distribution of pore sizes on a monolith of silica aerogel, and a type of integral plastic. These materials have been created by strictly controlling the vaporization of the volatile during the production process^{11,12}. Gradients in chemical composition have been achieved, for example: (1) chemical vapour deposition^{13,14}, (2) powder methods such as slip cast or dry processing¹⁵, (3) various coating methods¹⁶, and (4) thermal chemical reaction^{2,17}. Of these, (1) and (4) are relatively expensive, complicated and result in damage to bulk substrates. (2) and (3) produce stepped gradient structures, and it is difficult to control the thickness of each layer to less than 100 nm. Furthermore, most of these processes are not easily adapted to coating samples in the form of fibre bundles, fine powders or other materials with complicated shapes.

We have addressed the issue of establishing an inexpensive and widely applicable process for creating a material with a compositional gradient and excellent functionality. A schematic representative of our new *in situ* formation process for functional surface layers, which have a gradient-like structure towards the surface, is shown in Fig. 1. The important feature of our method is that the surface layer of the ceramic is not deposited on the substrate, but is formed during the production of the bulk ceramic. We confirmed that our process is applicable to any type of system as long as, in the green-body (that is, not-calcined) state, the system contains a resin and a low-molecular-mass additive that can be converted into a functional ceramic at high temperatures. Here, the resin is a type of precursor polymer (polycarbosilane, polycarbosilazane, polysilastylene, methylchloropolysilane, and so on) or binder polymer used for preparing green bodies from ceramic powders¹⁸. Although the former case (using precursor polymers) is explained in detail in this Letter, the latter case using binder polymers was also confirmed by treating a Si₃N₄ body with a TiN surface layer. Si₃N₄ can exhibit excellent thermal stability and wear resistance in the high-speed machining of cast iron, but shows poor chemical wear resistance in the machining of steel¹⁹. In order to avoid this problem, TiN coating, by means of expensive chemical vapour deposition, has often been performed on previously prepared Si₃N₄ substrates. But if our process is appropriately applied, formation of the TiN surface layer could be achieved during the sintering process of the Si₃N₄ green body. In this case, titanium(IV) butoxide and polystyrene are used as the low-molecular-mass additive and binder polymer, respectively. By a combination of sufficient maturation (in air at 100 °C) and subsequent sintering (in NH₃+H₂+N₂ at 1,200 °C), Si₃N₄ covered with TiN is successfully produced. This technology would be very useful for producing ceramic materials with complicated shapes and various coating layers. Moreover, our process is advantageous for preparing precursor ceramics (particularly fine particles, thin fibrous ceramics and films). The systems to which our concept is applicable are shown in Fig. 1.

Here we give a detailed account of the results for the precursor ceramic obtained using polycarbosilane. Polycarbosilane (–SiH(CH₃)–CH₂–)_n is a representative pre-ceramic polymer for preparing SiC ceramics—for example, Hi-Nicalon fibre²⁰ and Tyranno SA fibre²¹. Furthermore, oxide or nitride can also be produced from the polycarbosilane by firing in air or ammonia, respectively. Our new technology makes full use of the bleed-out

phenomenon^{5–9} of additives intentionally mixed in the polycarbosilane. Here, we have treated a polycarbosilane with $\text{Ti}(\text{OC}_4\text{H}_9)_4$ or $\text{Zr}(\text{OC}_4\text{H}_9)_4$, and have created a strong, fibrous photocatalyst with a surface TiO_2 layer, or a highly alkali-resistant SiC -based fibre with a surface ZrO_2 layer.

The pre-ceramic polymer containing $\text{Ti}(\text{OC}_4\text{H}_9)_4$ (50 wt%) was shaped into a fibre by melt spinning. The fibre was then matured in air at 70 °C for 100 h; during this process, the titanium compound oozed from the pre-ceramic polymer. The bleed-out phenomenon was confirmed using electron spectroscopy for chemical analysis (ESCA). After the maturation, the fibre material was cured in air at 200 °C. The cured material was then calcined in air. Figure 2a shows Auger electron spectroscopy (AES) depth analysis of the surface layer of the fibre, which was calcined at 1,200 °C in air after the maturation. The thickness of the surface layer can be controlled in the range 5–500 nm by changing the temperature (70–100 °C) and time (10–200 h) of the maturation²². Furthermore, the composition of the surface layer reflects the composition of the low-molecular mass additives.

According to the results of X-ray diffraction analysis, these fibres were composed of anatase- TiO_2 (crystalline size, ~8 nm) and an amorphous SiO_2 phase. Investigation by scanning electron microscopy (SEM) showed that the outer surface of the fibre prepared via maturation for 100 h appeared smooth and almost featureless (Fig. 3a). However, the transmission electron microscopy (TEM) image of the cross-section near the surface showed a TiO_2 -sintered surface and particle-dispersed bulk structures (Fig. 3b). Most of the surface TiO_2 crystals were directly sintered simultaneously (directly sintered structure), whereas the internal TiO_2 crystals were bound with the amorphous SiO_2 phase (liquid phase sintered structure). This result corresponds to the gradient composition shown in the AES data. The gradient-like structure resulted in strong adhesion between the surface TiO_2 layer and the bulk material, which is different from the behaviour of other coating layers formed on substrates by means of conventional methods. Although in the case of the other TiO_2 -covered silica fibre the TiO_2 layer was easily peeled

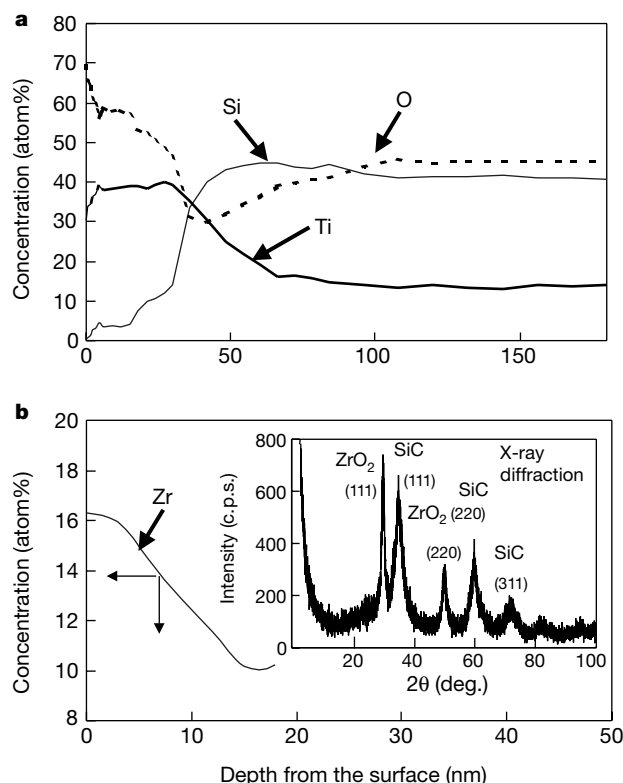


Figure 2 Surface gradient compositions and crystalline structure. **a**, Auger electron spectroscopy (AES) depth profiles of Ti, Si and O near the surface of the oxide fibre prepared from polycarbosilane containing $\text{Ti}(\text{OC}_4\text{H}_9)_4$. **b**, AES depth profile of Zr near the surface of the SiC fibre prepared from polycarbosilane containing $\text{Zr}(\text{OC}_4\text{H}_9)_4$, along with its X-ray diffraction pattern. The depth profiles were obtained using an ULVAC PHI SMART-200 operating at 3 kV. The X-ray diffraction pattern was recorded using a Rigaku X-ray diffractometer with $\text{CuK}\alpha$ radiation with a nickel filter.

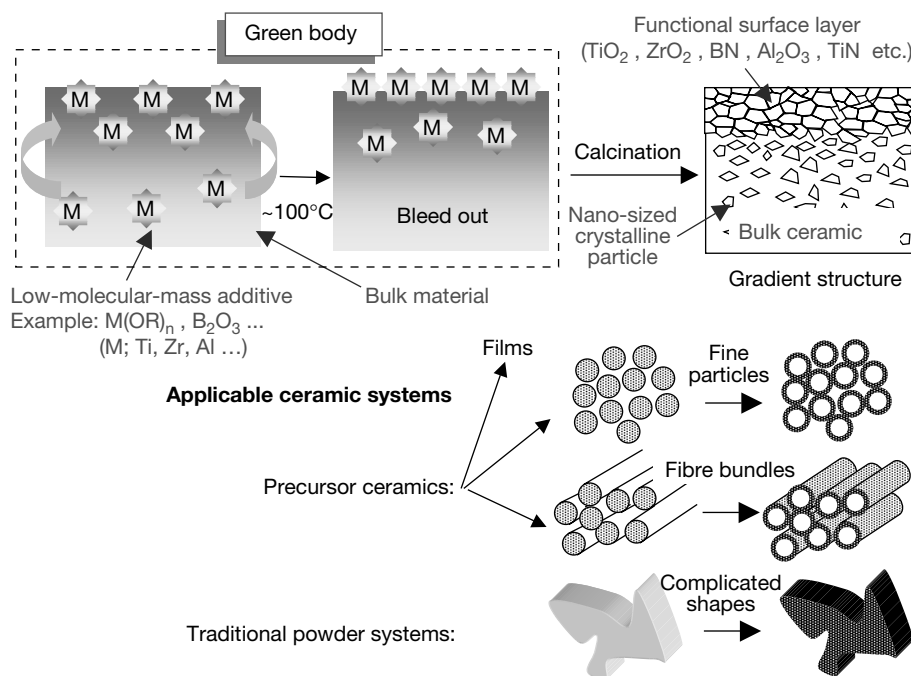


Figure 1 Schematic diagram of a general process for *in situ* formation of functional surface layers on ceramics. 'Precursor ceramics' indicates precursor polymers (polycarbosilane, polysilazone, polysilastyrene, methylchloropolysilane, and so on),

including low-molecular-mass additives. 'Traditional powder systems' indicate green bodies composed of ceramic powders and resins including low-molecular-mass additives.

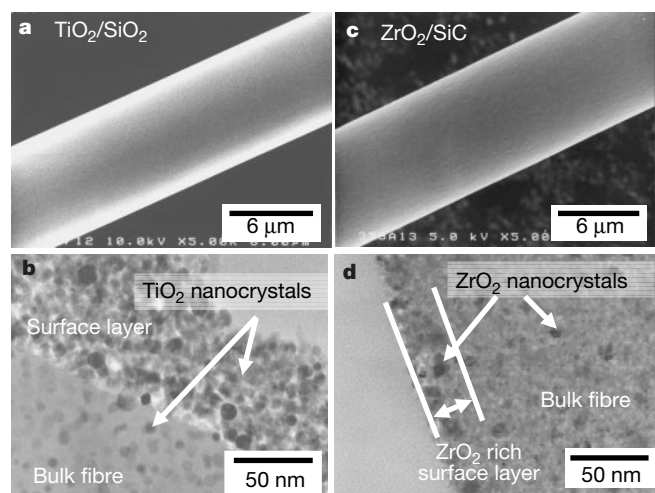


Figure 3 Surface appearances and cross-sections of fibre materials. SEM micrographs and TEM images of $\text{TiO}_2/\text{SiO}_2$ fibre (**a**, **b**) and ZrO_2/SiC fibre (**c**, **d**). **a** and **c**, SEM micrographs of the fibre surface. **b** and **d** TEM images of the cross-section near the fibre surface. SEM micrographs and TEM images were obtained using a Hitachi S-5000 operating at 5 kV and a JEM 2010F operating at 200 kV, respectively.

off, our TiO_2 -sintered surface definitely did not drop off after heat-cycling, washing or rubbing.

Furthermore, a strengthening effect of the fine particles (<10 nm) dispersed in the bulk ceramics can be achieved at the same time using our technology. Tensile testing of the monofilaments according to the ASTM D3379-75 standard with a 25-mm gauge length demonstrated that the strength was markedly higher (>2.5 GPa) than that of ordinary sol-gel $\text{TiO}_2/\text{SiO}_2$ fibres (<1 GPa)²³. At present, the reason for this is not clear, but it was assumed that this was caused by less stress-concentration at the surface region because of the gradient structure towards the surface.

We subsequently confirmed the objective function (photocatalytic activity) of the above-mentioned fibre as follows. We prepared a quartz tubular reactor (inner volume: 7 cm^3) with an ultraviolet (UV) lamp. The tubular reactor was filled with 2 g of our fibre material. The photocatalytic activity was confirmed using air containing 50 p.p.m. of acetaldehyde, at a flow rate of 0.2 l min^{-1} , and UV light at an intensity of 3 mW cm^{-2} (wavelength, 352 nm). Under these conditions (single pass), the acetaldehyde was completely decomposed accompanied by the generation of CO_2 . We also confirmed the coliform-sterilization ability of this fibre as follows. Our fibre (0.2 g) was placed in wastewater (20 ml) containing coliform at a concentration of $2 \times 10^6\text{ ml}^{-1}$. Irradiation by UV light (wavelength; 352 nm, 2 mW cm^{-2}) was performed at room temperature, and a small amount of the wastewater was extracted every hour. After cultivation using the extracted water, the amount of active coliform was calculated from the number of colonies formed. In this experiment, all of the coliform in the wastewater was completely sterilized within three hours. In the comparative study (with no fibre), the coliform content markedly increased to $3 \times 10^7\text{ ml}^{-1}$ under the same conditions.

The next fibre was prepared from polycarbosilane containing $\text{Zr}(\text{OC}_4\text{H}_9)_4$ by the same process as that used for the $\text{TiO}_2/\text{SiO}_2$ fibre material, except that the calcination was performed in Ar atmosphere at $1,400^\circ\text{C}$. In this case, the polycarbosilane and $\text{Zr}(\text{OC}_4\text{H}_9)_4$ were effectively converted into SiC-based bulk ceramic and zirconium oxide (cubic zirconia), respectively (X-ray diffraction results are given in Fig. 2b). Before the conversion, bleed-out of the zirconium compound proceeded effectively. AES depth analysis of the fibre surface showed an increase in the concentration of zirconium towards the surface (Fig. 2b). This construction was

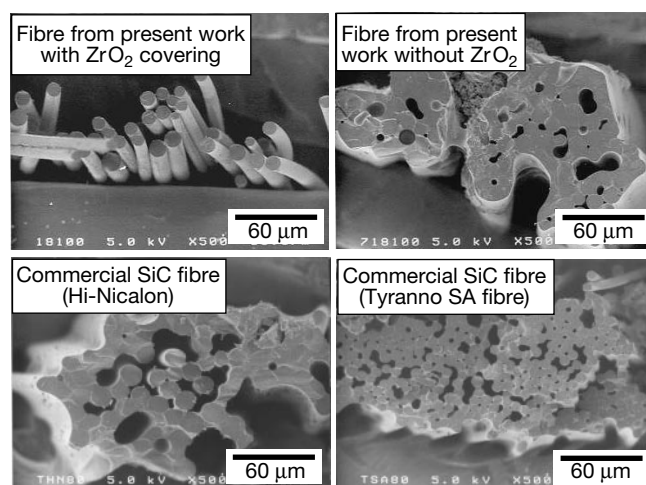


Figure 4 Alkali resistance of our ZrO_2/SiC fibre with comparative results. Shown are SEM micrographs of each SiC-based fibre, which had been immersed in deionized water saturated with potassium acetate and then annealed at 800°C for 100 h.

confirmed by the TEM image of the cross-section near the fibre surface (Fig. 3c and d). This indicates the direct production of an SiC-based fibre covered with a ZrO_2 surface layer, which has a gradient-like composition towards the surface. In general, amorphous fibres covered with ceramic crystal do not show high strength²⁴. However, this fibre showed relatively high strength (2.5 GPa) compared with other SiC fibre (2.1 GPa) coated with zirconia nanocrystals by means of the sol-gel method. The initial strength of the SiC fibre used for the comparative study was 3.1 GPa. The ZrO_2 surface layer, a basic oxide material, can provide better alkali resistance for SiC ceramics.

In order to confirm the better alkali resistance for our ZrO_2 -covered SiC fibre, the following experiment was performed. The fibre material was immersed for 15 min in deionized water saturated with potassium acetate and then annealed at 800°C for 100 h in air after drying. Comparative studies were conducted using the SiC-based fibre prepared from polycarbosilane, which did not contain zirconium (IV) butoxide, as well as commercial SiC fibres, namely, Hi-Nicalon and an alkali-resistant sintered SiC fibre²¹ (Tyranno SA fibre). Figure 4 shows the fractured surfaces of the tested fibre bundles, obtained using field-emission scanning electron microscopy (FE-SEM). As can be seen from the micrographs, only the ZrO_2 -covered SiC fibre, which was obtained using our method, retained its intact fibrous shape, whereas the other SiC fibres were extensively oxidized and then bonded together.

Fine powders (SiC , Si_3N_4 , SiO_2) covered with functional layers (BN, TiN, TiO_2 , ZrO_2 , Al_2O_3) could also be synthesized by sufficiently maturing and firing the precursor powders of polycarbosilane or polysilazane including the corresponding additives. The atmosphere and temperature during firing would need to be strictly controlled. □

Received 3 August; accepted 31 December 2001.

- Singh, R. K., Gilbert, D. R., Fitz-Gerald, J., Harkness, S. & Lee, D. G. Engineered interfaces for adherent diamond coatings on larger thermal-expansion coefficient mismatched substrate. *Science* **272**, 396–398 (1996).
- Gogotsi, Y. G. & Yoshimura, M. Formation of carbon films on carbides under hydrothermal conditions. *Nature* **367**, 628–630 (1994).
- Roy, R. Ceramics by the solution-sol-gel route. *Science* **238**, 1664–1669 (1987).
- Besmann, T. M., Sheldon, B. W., Lowden, R. A. & Stinton, D. P. Vapor-phase fabrication and properties of continuous-filament ceramic composites. *Science* **253**, 1104–1109 (1991).
- Kerk, S. L., Tay, S. C. & Hu, S. J. Effect of inherent deformations of leadframes on bleedability of plastic dip. *J. Electron. Mater.* **18** Part 1, 117–121 (1989).
- Petrovic, A. Morphological instability of poly(ethylene terephthalate) cyclic oligomer crystals. *J. Mater. Sci.* **20**, 1370–1374 (1985).

7. Perovic, A. & Murti, D. K. The effect of coatings on the surface precipitation of oligomeric crystals in poly(ethylene terephthalate) films. *J. Appl. Polym. Sci.* **29** Part 2, 4321–4327 (1984).
8. Needles, H. L., Berns, R. S., Lu, W. C., Alger, K. & Varma, D. S. Effect of nonionic surfactant and heat on selected properties on polyester. *J. Appl. Polym. Sci.* **25**, 1737–1744 (1980).
9. Perovic, A. & Sundararajan, P. R. Crystallization of cyclic oligomers in commercial poly(ethylene terephthalate) films. *Polym. Bull.* **6**, 277–283 (1982).
10. Neubrand, A. & Rodel, J. Gradient materials: An overview of a novel concept. *Z. Metallkd.* **88**, 358–371 (1997).
11. Gladden, L. F. *et al.* Structural studies of high surface area silicas. *J. Non-Cryst. Solids* **139**, 47–59 (1992).
12. Dariel, M. P., Levin, L. & Frage, N. Graded ceramic preforms: Various proceeding approaches. *Mater. Chem. Phys.* **67**, 192–198 (2001).
13. Fischbach, D. B. & Lemoine, P. M. Influence of a CVD carbon coating on the mechanical property stability of nicalon silicon carbide fibre. *Composites Sci. Technol.* **37**, 55–61 (1990).
14. Li, J. X., Matsuo, Y. & Kimura, S. Improvement of thermal stability of SiC fibre by CVD-C/SiC coating. *Bull. Ceram. Soc. Jpn* **99**, 1207–1211 (1991).
15. Moya, J. S., Sanchez-Herencia, M. R., Pena, P. & Requena, J. Layered ceramic composites: A new family of advanced materials. *Third Euro-Ceramics* (eds Durán, P. & Fernández, J. F.) **3**, 289–300 (Feenza Editrice Iberica, Spain, 1993).
16. Sidky, P. S. & Hocking, M. G. Review of inorganic coatings and coating processes for reducing wear and corrosion. *Br. Corrosion J.* **34**, 171–183 (1999).
17. Somya, S. (ed.) *Hydrothermal Reactions for Materials Science and Engineering* (Elsevier, London, 1989).
18. Varaprasad, D. V., Wade, B., Venkatasubramanian, N., Desai, P. & Abhiraman, A. S. Critical requirements in the formation of continuous ceramic fibre precursors. *Ind. J. Fibre Textile Res.* **16**, 73–82 (1991).
19. Bhat, D. G., Rebenne, H. E. & Strandberg, C. Analysis of coating interlayer between silicon nitride cutting tools and titanium carbide and titanium nitride coatings. *J. Mater. Sci.* **26**, 4567–4580 (1991).
20. Takeda, M., Sakamoto, J., Saeiki, A., Imai, Y. & Ichikawa, H. High performance silicon carbide fibre Hi-Nicalon for ceramic matrix composites. *Ceram. Eng. Sci. Proc.* **16**, 37–44 (1995).
21. Ishikawa, T., Kohtoku, Y., Kumagawa, K., Yamamura, T. & Nagasawa, T. High-strength alkali-resistant sintered SiC fibre stable to 2,200 °C. *Nature* **391**, 773–775 (1998).
22. Yamaoka, H., Harada, Y. & Fujii, T. Inorganic fibre including zirconia and its production process. Japanese patent Application No. 2001-167679 (2001).
23. Abe, Y., Gunji, T. & Hikita, M. Preparation of SiO₂-TiO₂ fibres from polytitanosiloxanes. *Yogyokokai-shi* **94** (12), 1243–1245 (1986).
24. Shen, L., Tan, B. J., Willis, W. S., Galasso, F. S. & Suib, S. L. Characterization of dip-coated boron nitride on silicon carbide fibres. *J. Am. Ceram. Soc.* **77**, 1011–1016 (1994).

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to T.I.
(e-mail: 24613u@ube-ind.co.jp).

Observation and interpretation of a time-delayed mechanism in the hydrogen exchange reaction

Stuart C. Althorpe^{*†}, Félix Fernández-Alonso[‡], Brian D. Bean[§],
James D. Ayers[§], Andrew E. Pomerantz[§], Richard N. Zare[§]
& Eckart Wrede^{*}

^{*} Department of Chemistry, University of Durham, Durham DH1 3LE, UK

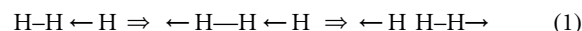
[‡] Istituto di Struttura della Materia – Consiglio Nazionale delle Ricerche,
Area della Ricerca di Roma – Tor Vergata, via del Fosso del Cavaliere 100,
00133 Rome, Italy

[§] Department of Chemistry, Stanford University, Stanford,
California 94305-5080, USA

Extensive theoretical^{1–13} and experimental^{2,13–22} studies have shown the hydrogen exchange reaction $\text{H} + \text{H}_2 \rightarrow \text{H}_2 + \text{H}$ to occur predominantly through a ‘direct recoil’ mechanism: the H–H bonds break and form concertedly while the system passes straight over a collinear transition state, with recoil from the collision causing the H₂ product molecules to scatter backward. Theoretical predictions agree well with experimental observations of this scattering process^{15–20,22}. Indirect exchange mechanisms involving H₃ intermediates have been suggested to occur as

well^{8–13}, but these are difficult to test because bimolecular reactions cannot be studied by the femtosecond spectroscopies²³ used to monitor unimolecular reactions. Moreover, full quantum simulations of the time evolution of bimolecular reactions have not been performed. For the isotopic variant of the hydrogen exchange reaction, $\text{H} + \text{D}_2 \rightarrow \text{HD} + \text{D}$, forward scattering features²¹ observed in the product angular distribution have been attributed^{21,12} to possible scattering resonances associated with a quasibound collision complex. Here we extend these measurements to a wide range of collision energies and interpret the results using a full time-dependent quantum simulation of the reaction, thus showing that two different reaction mechanisms modulate the measured product angular distribution features. One of the mechanisms is direct and leads to backward scattering, the other is indirect and leads to forward scattering after a delay of about 25 femtoseconds.

The dominant recoil mechanism in the $\text{H} + \text{H}_2$ reaction may be written as



where the arrows indicate the direction of motion of the species involved. This mechanism results in ‘backward scattered’ H₂ angular distributions, which recoil from the collision at centre-of-mass scattering angles θ close to 180° (where θ is measured from the approach direction of the H reagent at $\theta = 0^\circ$). A recent *ab initio* potential energy surface for this system, the BKMP2 surface⁴, is so accurate that the theoretical angular distributions, obtained by solving the exact time-independent Schrödinger equation for motion of the H atoms on this surface, are in excellent quantitative agreement with experiment^{15–20,22}.

Quasiclassical^{8,21} and quantum⁹ calculations have also predicted ridge structures along a line in the energy–angle (E – θ) plane which evolve into broad peaks in the ‘forward direction’ (around $\theta = 0^\circ$). The quasiclassical ridges were associated with a time-delayed (15–35 fs) mechanism, although this classical prediction should be interpreted with care because the quasiclassical method does not give an exact description of the scattering dynamics. The quasiclassical⁸ E – θ ridges and forward scattering features are less pronounced than their quantum⁹ counterparts, indicating that quantum effects are important in causing these phenomena. In the isotopic variant of the hydrogen exchange reaction, $\text{H} + \text{D}_2 \rightarrow \text{HD}(\nu' = 3, j' = 0) + \text{D}$, forward scattering²¹ has been observed at a collision energy $E = 1.64$ eV (ν' and j' refer to the vibrational and rotational quantum numbers of the diatomic product, respectively). For $\text{HD}(\nu' = 3, j' = 0)$, the quasiclassical calculations underestimated the amount of forward scattering by about a factor of three²¹, while model quantum calculations¹² indicated that a quantum threshold resonance, associated with a quasibound complex formed by the collision partners, was possible at this energy.

The experiments reported here used the PHOTOLoc (photo-initiated reaction analysed via the law of cosines) technique^{17,18,21,22}. A 1:9 mixture of HBr and D₂ is expanded in a free jet using a pulsed nozzle valve. The reaction is initiated by photolysing HBr with a 5-ns laser pulse, varying the photolysis wavelength between 203 and 225 nm to correspond to the nine collision energies shown in Fig. 1. The $\text{HD}(\nu' = 3, j' = 0)$ product is detected 15 ns after photoinitiation, using (2+1) resonance enhanced multiphoton ionization (REMPI) through the Q-branch members of the HD EF $1^1\Sigma_g^+ - X^1\Sigma_g^+$ (0,3) band²⁴. The energy resolution of the experiment is about 50 meV full-width at half-maximum, due mainly to the residual translational temperature ($T \approx 50$ K) of the beam expansion.

A direct comparison of the experimental measurements with the theoretical predictions is presented in Fig. 1 in the form of time-of-flight (TOF) profiles of the $\text{HD}(\nu' = 3, j' = 0)$ product. The theoretical TOF profiles were calculated by solving the exact time-dependent Schrödinger equation for the $\text{H} + \text{D}_2(\nu = 0, j = 0)$ reaction on the BKMP2 potential energy surface⁴, using a recently developed

[†] Present address: School of Chemistry, University of Exeter, Stocker Road, Exeter EX4 4QD, UK.

wavepacket method²⁵, extended as explained below. Angular distributions, at fixed values of E from 1.3 to 2.2 eV, were calculated by summing all partial waves from total angular momentum $J = 0$ to 30, and were converged to better than 2%. The convergence of selected fixed- J reaction probabilities was checked by comparing with results calculated using the CCP6 time-independent code²⁶. The calculated angular distributions were then convoluted with the instrumental and collision-energy resolution of the experiment to obtain the TOF profiles shown in Fig. 1.

The experimental and theoretical TOF profiles in Fig. 1 are in excellent agreement. The most notable feature in the TOF profiles is the overall change in shape as a function of collision energy, which we interpret by referring to the calculated HD($v' = 3, j' = 0$) angular distributions as a function of collision energy in Fig. 2. (Strictly speaking Fig. 2 shows the angular distributions multiplied by a factor of $\sin\theta$, so as to give a fair representation of the amount of reaction product per solid angle.) On passing from a collision energy of 1.5 eV to 1.8 eV, the angular distribution changes from being mainly sideways scattered to being (asymmetrically) forward-backward peaked. This change maps onto the TOF profiles as a change from a broad pair of peaks around ± 25 ns, at 1.5 eV, to three narrow peaks around 0 ns and ± 50 ns, at 1.8 eV. The ± 25 -ns peaks correspond to sideways scattering in Fig. 2, the 0-ns peak to back scattering, and the ± 50 -ns side peaks to forward scattering. Note that the reduction in height of the ± 50 -ns peaks between 1.64 and 1.85 eV is caused by a reduction in experimental sensitivity to forward-scattering features—not by a reduction in the forward scattering itself. The forward scattering remains strong between 1.64 and 1.85 eV while becoming narrower. This feature is the onset of a

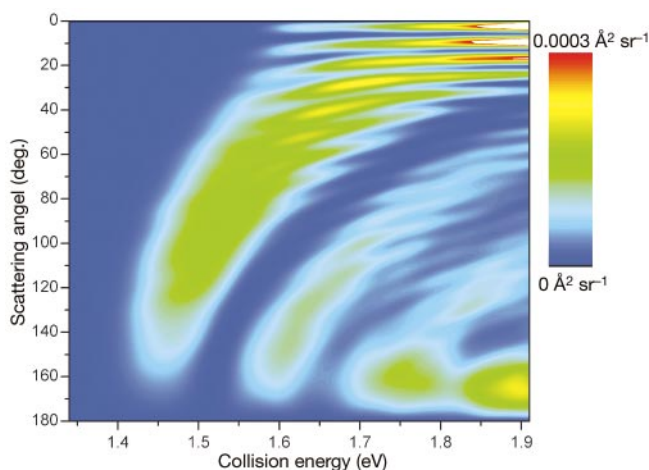


Figure 2 Calculated HD($v' = 3, j' = 0$) angular distributions. The distributions [$d\sigma(E, \theta)/d\Omega \sin\theta$] are plotted as a function of collision energy E and centre-of-mass scattering angle θ . The forward ($\theta \approx 0^\circ$) peak at $E > 1.6$ eV shows up as the side-peaks at ± 50 ns in the TOF profiles of Fig. 1. The ridge between collision energies of 1.4 and 1.7 eV causes the broadening of the TOF profiles between collision energies of 1.39 to 1.64 eV.

broad forward scattering peak as a function of collision energy, such as predicted previously for the H + H₂ (ref. 9) and D + H₂ (refs 8 and 9) reactions at lower collision energies. Our TOF profiles are, to our knowledge, the first experimental measurement of this feature, and also of the E - θ 'ridge', which, for H + D₂ $\rightarrow$ HD($v' = 3, j' = 0$) + D,

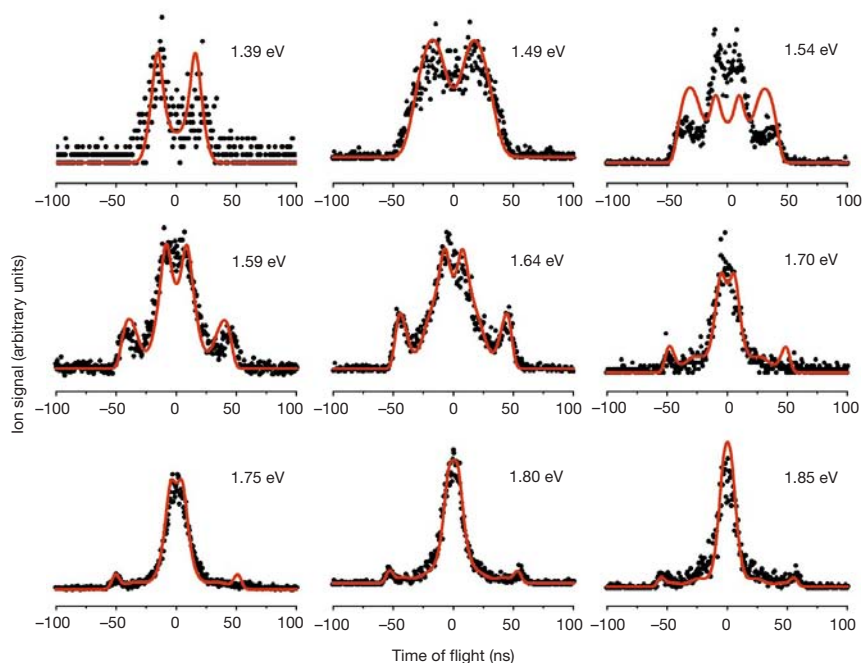


Figure 1 Comparison between experimental and theoretical results for the H + D₂ $\rightarrow$ HD($v' = 3, j' = 0$) + D reaction. Each frame is a plot of an experimental (black dots) and theoretical (red curve) time-of-flight (TOF) profile, showing the number of ionized HD($v' = 3, j' = 0$) molecules that arrive at the detector as a function of time. There is a one-to-one mapping between this arrival time (which should not be confused with the time t used to follow the reaction mechanism in Fig. 3) and the centre-of-mass scattering angle θ (measured from the approach direction of the H reagent at $\theta = 0^\circ$). Each black dot represents the total HD($v' = 3, j' = 0$) single-ion count at a given time of arrival at the detector. The spread in the experimental points is indicative of the statistical errors arising from the Poisson-like statistics of the ion-counted signals. Systematic errors were identified and reduced to a minimum by performing at least three independent

measurements of the same TOF profile at each collision energy. The TOF profiles are symmetric about the centre because there are as many reagent H atoms with speed projections towards the detector as those away from the detector. The theoretical TOF profiles were obtained by convoluting the calculated angular distributions of Fig. 2 with the instrumental and collision-energy resolution of the experiment. At present we do not understand the cause of the discrepancies at 1.54 eV, although we note that it is around this energy that the shape of the angular distributions changes most rapidly as a function of collision energy; also, the calculations included only the lowest rotational level of the D₂($v = 0$) reagent, whereas higher rotational levels ($j < 3$) contribute to the experimental signal. The excellent overall agreement between theory and experiment means that the simulation shown in Fig. 3 is a realistic picture of the reaction mechanism.

lies between 1.4 and 1.7 eV (see Fig. 2). As E increases from 1.4 to 1.7 eV, the ridge moves in the forward direction, evolving into the forward peak. Hence the scattering features predicted previously^{8,9} and attributed to indirect processes are responsible for the overall change in the shape of the TOF profiles.

The excellent agreement between experiment and theory suggests that the wavefunction we use provides a realistic description of the motion of the atoms during the reaction. To obtain the time-dependent wavefunction, we built on previous work^{27–29} on calculating time-dependent wavefunctions for fixed values of J (usually $J = 0$). We extend the method of ref. 25 to calculate the fixed- J wavefunctions for every value of J that contributes to the angular distributions (that is, $J = 0–30$). The fixed- J wavefunctions are then superposed to give the complete wavefunction, Ψ , which describes the motion of the nuclei as a function of time, t , and of the six coordinates that locate the nuclei in the centre-of-mass frame. We next project Ψ onto the HD($v' = 3, j' = 0$) rovibrational product state, obtaining the function $\Phi(R, \theta, t)$ which describes when and where the HD($v' = 3, j' = 0$) product is formed, and how it scatters. Note that R is the distance from the product HD centre of mass to the product D atom, and θ is the centre-of-mass scattering angle defined above.

We calculate $|\Phi(R, \theta, t)|^2 \sin \theta$ at 120 values of t from 0 to 100 fs. Plots at selected values of t are shown in blue in Fig. 3. The

symmetry about the H–DD approach axis reflects the cylindrical symmetry of the reaction in three-dimensional space, which is ensured at $t = 0$ by the symmetry of the $D_2(v = 0, j = 0)$ wavefunction, and is preserved throughout the collision. The spatial part of Ψ at time $t = 0$ is a plane-wave wavepacket $\chi(R', \theta')$, which describes the H atom approaching $D_2(v = 0, j = 0)$ from a mean separation R' of 6 bohr (where (R', θ') are the reagent analogues to (R, θ)). The function $|\chi(R', \theta')|^2$ is plotted in green in Fig. 3. Note that $\chi(R', \theta')$ is made up of a superposition of time-independent wavefunctions, describing the H–DD approach motion over a continuous range of collision energies from 1.33 to 2.2 eV. The distribution of these energies is flat, damping to zero at the edges. (The precise form of the distribution is a Fourier-transformed distributed approximating functional³⁰, centred about 1.65 eV, with σ and M (defined in ref. 30) set to 0.07 eV and 88.) Using different energy distributions confirmed that all the key features in Fig. 3 were independent of the form of the energy distribution.

Figure 3 demonstrates that the HD($v' = 3, j' = 0$) product is formed by two distinct reaction mechanisms. The first mechanism is the recoil mechanism of equation (1), which ejects the HD product in the opposite direction to the H–DD approach direction. The second mechanism occurs roughly 25 fs after the recoil mechanism, and ejects most of the HD product in the opposite (forward) direction. This time-delayed mechanism causes the broad

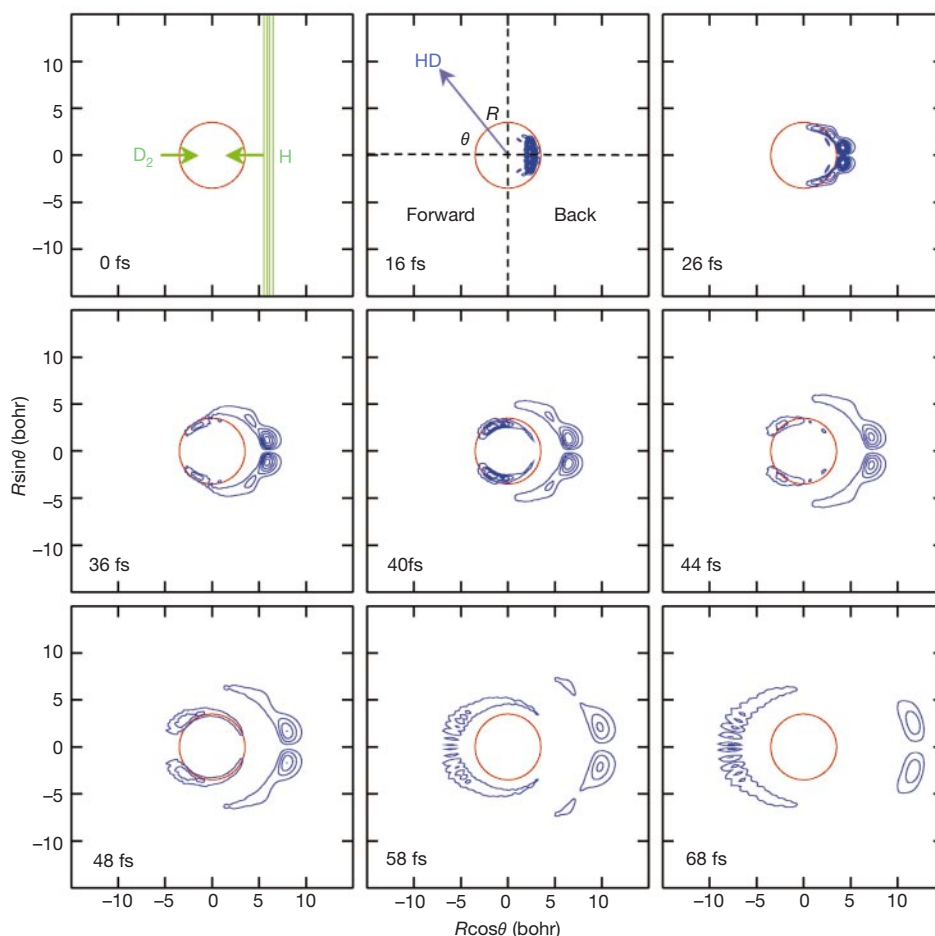


Figure 3 Snapshots from the quantum simulation of the $H + D_2(v = 0, j = 0) \rightarrow HD(v' = 3, j' = 0) + D$ reaction. The blue contours are obtained from the complete wavefunction of the reaction as explained in the text, and show the time-evolution of the HD($v' = 3, j' = 0$) product, as a function of the centre-of-mass scattering angles R and θ . The green contours and arrows show the location and direction of travel of the reagents at $t = 0$. The red circles are of radius $R = 3.5$ bohr and give a rough indication of the extent of the transition state region. Two reaction mechanisms are visible, separated by a time delay of

about 25 fs. The earlier mechanism, which throws out most of its products in the backward direction, is the direct 'recoil' mechanism, and is responsible for the central peaks in the TOF profiles of Fig. 1. The time-delayed mechanism, which throws out most of its products in the forward direction, is responsible for the side peaks in the TOF profiles of Fig. 1. Snapshots are shown here for selected values of t ; the entire sequence of frames is available as a movie (at http://www.dur.ac.uk/chemistry/publications/sc_althorpe/nature.html).

forward peak in Fig. 2, and hence the side peaks in the TOF profiles in Fig. 1. Analysis of the angular distributions showed that the time-delayed mechanism is also mainly responsible for the sideways ridge (see Fig. 2), and hence for the broadening of the TOF profiles around $E = 1.4$ – 1.7 eV.

We see that the HD product from the time-delayed mechanism starts to form at scattering angles of roughly 45° , where the wavefunction is found to have mostly 'high impact parameter' components, corresponding to $J = 15$ – 20 . This indicates that the H atoms that react via this mechanism come from the parts of the plane wave (green function in Fig. 3) around $R\sin\theta = \pm(1.1$ – $1.6)$ bohr. It is likely therefore that H attacks the D_2 at a glancing angle, in contrast to the direct mechanism (in which H attacks the D_2 bond collinearly). If this happens, then energy initially concentrated in the HDD bend must be converted into energy along the HDD asymmetric stretch, to allow the system to pass over the transition state and react.

The successful combination of experiment and theory to elucidate in detail the mechanisms involved in one of the simplest chemical reactions, the hydrogen exchange reaction, can also be applied to more complex reactions. These include the reactions $O(^1D) + H_2 \rightarrow OH + H$, $F + HD \rightarrow HF + D$, and $H + H_2O \rightarrow OH + H_2$, for which experimental angular distributions and realistic potential energy surfaces are available^{19,20}, and for which it is computationally feasible to solve the time-dependent Schrödinger equation. □

Received 24 October 2001; accepted 22 January 2002.

- Hirshfelder, J. O., Eyring, H. & Topley, B. Reactions involving hydrogen molecules and atoms. *J. Chem. Phys.* **4**, 170–177 (1936).
- Truhlar, D. G. & Wyatt, R. E. History of H_3 kinetics. *Annu. Rev. Phys. Chem.* **27**, 1–43 (1976).
- Neuhauser, D. et al. State-to-state rates for the $D + H_2(v = 1, j = 1) \rightarrow HD(v', j') + H$ reaction—predictions and measurements. *Science* **257**, 519–522 (1992).
- Boothroyd, A. I., Keogh, W. J., Martin, P. G. & Peterson, M. R. A refined H_3 potential surface. *J. Chem. Phys.* **104**, 7139–7152 (1996).
- Zhang, J. Z. H. & Miller, W. H. Quantum reactive scattering via the S-matrix version of the Kohn variational principle: integral cross sections for $H + H_2(v_1 = j_1 = 0) \rightarrow H_2(v_2 = 1j_2 = 1, 3) + H$ in the energy range $E_{\text{total}} = 0.9$ – 1.4 eV. *Chem. Phys. Lett.* **153**, 465–470 (1988).
- Miller, W. H. Recent advances in quantum mechanical reactive scattering theory. *Annu. Rev. Phys. Chem.* **41**, 245–281 (1990).
- D'Mello, M., Manolopoulos, D. E. & Wyatt, R. E. Quantum dynamics of the $H + D_2 \rightarrow D + HD$ reaction: comparison with experiment. *J. Chem. Phys.* **94**, 5985–5993 (1991).
- Aoiz, F. J., Herrero, V. J. & Sáez Rábanos, V. Quasiclassical state to state reaction cross sections for $D + H_2(v = 0, j = 0) \rightarrow HD(v', j') + H$. Formation and characteristics of short-lived collision complexes. *J. Chem. Phys.* **97**, 7423–7436 (1992).
- Miller, W. H. & Zhang, J. Z. H. How to observe the elusive resonances in H or $D + H_2 \rightarrow H_2$ or $HD + H$ reactive scattering. *J. Phys. Chem.* **95**, 12–19 (1991).
- Schatz, G. C. & Kuppermann, A. Dynamical resonances in collinear, coplanar, and three-dimensional quantum mechanical reactive scattering. *Phys. Rev. Lett.* **35**, 1266–1269 (1973).
- Skodje, R. T., Sadeghi, R., Köppel, H. & Krause, J. L. Spectral quantization of transition state dynamics for the three-dimensional $H + H_2$ reaction. *J. Chem. Phys.* **101**, 1725–1729 (1994).
- Allison, T. C., Friedman, R. S., Kaufman, D. J. & Truhlar, D. G. Analysis of the resonance in $H + D_2 \rightarrow HD(v' = 3) + D$. *Chem. Phys. Lett.* **327**, 439–445 (2000).
- Fernández-Alonso, F. & Zare, R. N. Scattering resonances in the simplest chemical reaction. *Annu. Rev. Phys. Chem.* **53**, 67–99 (2002).
- Nieh, J.-C. & Valentini, J. J. Experimental observation of dynamical resonances in the $H + H_2$ reaction. *Phys. Rev. Lett.* **60**, 519–522 (1988).
- Schneider, L., Seekamp-Rahn, K., Wrede, E. & Welge, K. H. Experimental determination of quantum state resolved differential cross sections for the hydrogen exchange reaction $H + D_2 \rightarrow HD + D$. *J. Chem. Phys.* **107**, 6175–6195 (1997).
- Wrede, E. et al. The dynamics of the hydrogen exchange reaction at 2.20 eV collision energy: comparison of experimental and theoretical differential cross sections. *J. Chem. Phys.* **110**, 9971–9981 (1999).
- Fernández-Alonso, F., Bean, B. D. & Zare, R. N. Differential cross sections for $H + D_2 \rightarrow HD(v' = 1, j' = 1, 5, 8) + D$ at 1.7 eV. *J. Chem. Phys.* **111**, 1035–1042 (1999).
- Fernández-Alonso, F., Bean, B. D. & Zare, R. N. Differential cross sections for $H + D_2 \rightarrow HD(v' = 2, j' = 0, 3, 5) + D$ at 1.55 eV. *J. Chem. Phys.* **111**, 2490–2498 (1999).
- Liu, K. Crossed-beam studies of neutral reactions: state-specific differential cross sections. *Annu. Rev. Phys. Chem.* **52**, 139–164 (2001).
- Casavecchia, P. Chemical reaction dynamics with molecular beams. *Rep. Prog. Phys.* **63**, 355–414 (2000).
- Fernández-Alonso, F. et al. Evidence for scattering resonances in the $H + D_2$ reaction. *Angew. Chem. Int. Edn Engl.* **39**, 2748–2752 (2000).
- Fernández-Alonso, F. et al. Forward scattering in the $H + D_2 \rightarrow HD + D$ reaction: comparison between experiment and theoretical predictions. *J. Chem. Phys.* **115**, 4534–4545 (2001).
- Zewail, A. H. Femtochemistry: atomic-scale dynamics of the chemical bond. *J. Phys. Chem.* **104**, 5660–5694 (2000).

- Rinnen, K.-D., Buntine, M. A., Kliner, D. A. V., Zare, R. N. & Huo, W. M. Quantitative determination of H_2 , HD , and D_2 internal-state distributions by (2+1) resonance-enhanced multiphoton ionization. *J. Chem. Phys.* **95**, 214–225 (1991).
- Althorpe, S. C. Quantum wavepacket method for state-to-state reactive cross sections. *J. Chem. Phys.* **114**, 1601–1616 (2001).
- Skouteris, D., Castillo, J. F. & Manolopoulos, D. E. ABC: a quantum reactive scattering program. *Comput. Phys. Commun.* **133**, 128–135 (2000).
- Althorpe, S. C., Kouri, D. J. & Hoffman, D. K. Further partitioning of the reactant-product decoupling equations of state-to-state reactive scattering and their solution by the time-independent wavepacket method. *J. Chem. Phys.* **107**, 7816–7824 (1997).
- Judson, R. S., Kouri, D. J., Neuhauser, D. & Baer, M. Time-dependent wavepacket method for the complete determination of S-matrix elements for reactive molecular collisions in 3 dimensions. *Phys. Rev. A* **42**, 351–366 (1990).
- Balint-Kurti, G. G., González, A. I., Goldfield, E. M. & Gray, S. K. Quantum reactive scattering of $O(^1D) + H_2$ and $O(^1D) + HD$. *Faraday Discuss.* **110**, 169–183 (1998).
- Huang, Y., Zhu, W., Kouri, D. J. & Hoffman, D. K. Distributed approximating function approach to atom-diatom reactive scattering: Time-dependent and time-independent wavepacket treatments. *J. Phys. Chem.* **98**, 1868–1874 (1994).

Acknowledgements

We thank L. Bañares and J. F. Castillo for discussions on the $H + D_2$ reaction, and D. C. Clary, D. E. Manolopoulos and J. M. Hutson for reading the manuscript. S.C.A. thanks the Royal Society for a University Research Fellowship. F.F.A. acknowledges partial support by a European Union Marie Curie fellowship. The experimental part of this work was supported at Stanford by the US National Science Foundation.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.C.A. (e-mail: s.c.althorpe@ex.ac.uk).

Glacial–interglacial stability of ocean pH inferred from foraminifer dissolution rates

David M. Anderson* & David Archer†

* NOAA Paleoclimatology Program and University of Colorado, 325 Broadway, Boulder, Colorado 80303, USA

† Department of Geophysical Sciences, University of Chicago, 5734 S. Ellis Ave., Chicago, Illinois 60637, USA

The pH of the ocean is controlled by the chemistry of calcium carbonate. This system in turn plays a large role in regulating the CO_2 concentration of the atmosphere on timescales of thousands of years and longer. Reconstructions of ocean pH and carbonate-ion concentration are therefore needed to understand the ocean's role in the global carbon cycle. During the Last Glacial Maximum (LGM), the pH of the whole ocean is thought to have been significantly more basic¹, as inferred from the isotopic composition of boron incorporated into calcium carbonate shells, which would partially explain the lower atmospheric CO_2 concentration at that time. Here we reconstruct carbonate-ion concentration—and hence pH—of the glacial oceans, using the extent of calcium carbonate dissolution observed in foraminifer faunal assemblages as compiled in the extensive global CLIMAP data set². We observe decreased carbonate-ion concentrations in the glacial Atlantic Ocean, by roughly $20 \mu\text{mol kg}^{-1}$, while little change occurred in the Indian and Pacific oceans relative to today. In the Pacific Ocean, a small ($5 \mu\text{mol kg}^{-1}$) increase occurred below 3,000 m. This rearrangement of ocean pH may be due to changing ocean circulation from glacial to present times, but overall we see no evidence for a shift in the whole-ocean pH as previously inferred from boron isotopes¹.

Foraminifer assemblage alteration by preferential dissolution has been recognized since the earliest studies in micropalaeontology. After the temperature of the surface environment, dissolution is the next most important variable in explaining the distribution of foraminifer faunas on the sea floor. Most foraminifer palaeotemperature equations identify a 'dissolution factor' that strongly correlates with water depth and isolates the variance due to dissolution as statistically independent from the variation due to sea surface temperature. Other factors include productivity³ and surface ocean mixed-layer depth⁴. The dominant species in the dissolution factor are typically *Globigerinoides sacculifer*, *Neoglobobulimina dutertrei*, *Pulleniatina obliquiloculata*, *Globorotalia menardii* and *Globorotalia tumida*⁵. The effect of carbonate ion saturation is apparent in plots of *N. dutertrei*, *Globigerinoides ruber* and *Globigerina bulloides* species abundance (%) versus carbonate-ion saturation and sea surface temperature (Fig. 1). In addition to the dominant correlation with temperature, these and other species have a less dramatic but statistically robust association with carbonate-ion content. The greatest abundance of *N. dutertrei* is found only in the environment where temperature is warm and the sea floor is undersaturated. Likewise, *G. bulloides*, most abundant in water of 8–12 °C, is less abundant in supersaturated environments and more abundant in undersaturated environments.

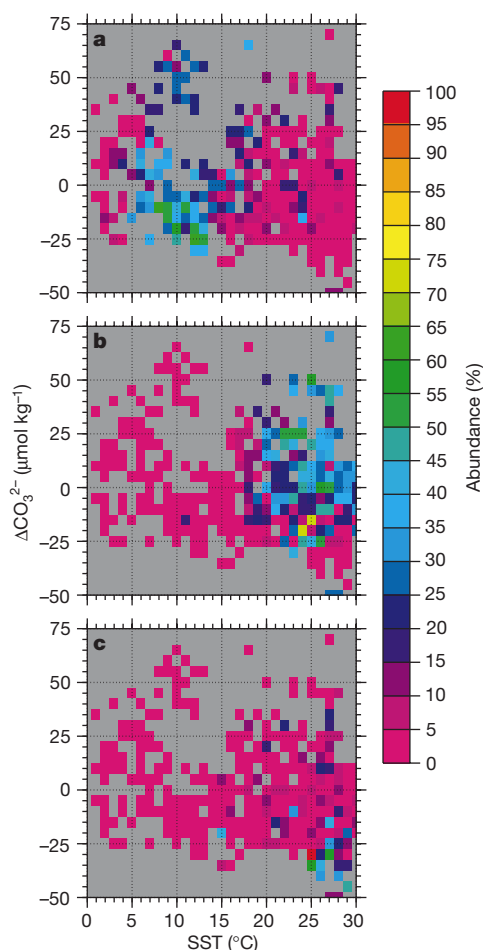


Figure 1 Foraminifer abundance in core-top sediments versus sea surface temperature and carbonate-ion saturation at the sea floor. Abundance (%) for three species (**a**, *Globigerina bulloides*; **b**, *Globigerinoides ruber*; and **c**, *Neoglobobulimina dutertrei*) from the core-top database of 905 samples are shown as a function of mean annual sea surface temperature (SST) and the carbonate-ion saturation (ΔCO_3^{2-}) corresponding to the water depth of the sample.

We reconstructed the saturation state of calcite, ΔCO_3^{2-} , of the glacial ocean from the global CLIMAP⁵ planktonic foraminifer abundance data set, using the modern analogue technique (see Methods). A data set of core-top species counts, coupled with modern deep-ocean ΔCO_3^{2-} values, provides the calibration. A test assemblage is compared with core-top data to find its five closest modern analogues. The mean of the analogue ΔCO_3^{2-} values provides the estimate of the test ΔCO_3^{2-} . The scatter in analogue ΔCO_3^{2-} provides an estimate of the uncertainty, which is generally 10 $\mu\text{mol kg}^{-1}$ or less. The same method has been used to estimate sea surface temperatures, to within 1.5 °C (ref. 5).

We calculated the absolute carbonate-ion concentration, $[\text{CO}_3^{2-}] = \Delta\text{CO}_3^{2-} + \text{CO}_{3\text{sat}}$ (refs 6, 7) and averaged the observations across 500-m depth intervals for the Atlantic ($n = 163$) and combined Indian and Pacific ($n = 108$) samples (Fig. 2). The data are shown in the context of calcite saturation (heavy solid line) and the mean $[\text{CO}_3^{2-}]$ and its standard deviation σ (light solid line and dashed envelope). The reconstructions reproduce both the mean and the scatter of the observed water column distributions of $[\text{CO}_3^{2-}]$. Each point is typically the average of at least ten samples (that is, fifty analogues) except for the two shallowest glacial Atlantic averages where $n = 4$.

The largest difference in the glacial ocean is in the Atlantic. Today freshly ventilated North Atlantic Deep Water (NADW), high in CO_3^{2-} , distinguishes the chemistry of the Atlantic from the Pacific and Indian oceans. According to palaeotracers ^{13}C (ref. 8) and Cd (ref. 9) there existed a glacial analogue of NADW (GNADW), which ventilated the water column to about 2.3-km depth^{10,11}. The intermediate Pacific^{12,13} and Indian¹⁴ oceans were also apparently more ventilated than today. In addition, the glacial decrease in atmos-

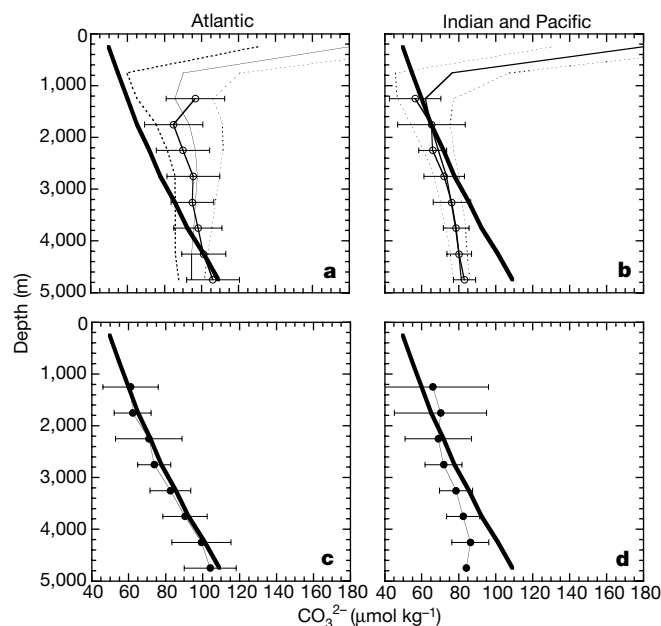


Figure 2 Vertical profiles of carbonate-ion concentration in the modern and glacial Atlantic and combined Indian and Pacific oceans. Modern estimates for the Atlantic (**a**) and Indian and Pacific (**b**) reconstructed from core-top samples ($n = 905$) averaged for 500-m depth intervals (open circles), are compared with the average (thin solid line) and range (dashed line, one standard deviation above and below the mean) in concentration from the modern gridded database, and with the saturation profile (thick line). Reconstructions for the Last Glacial Maximum for the Atlantic (**c**) ($n = 163$) and Indian and Pacific (**d**) ($n = 108$) averaged for 500-m depth intervals (filled circles) relative to saturation (thick line) for the Atlantic and the combined Indian and Pacific ocean samples. The uncertainty is the standard deviation of the CO_3^{2-} of all selected analogues for each 500-m depth interval.

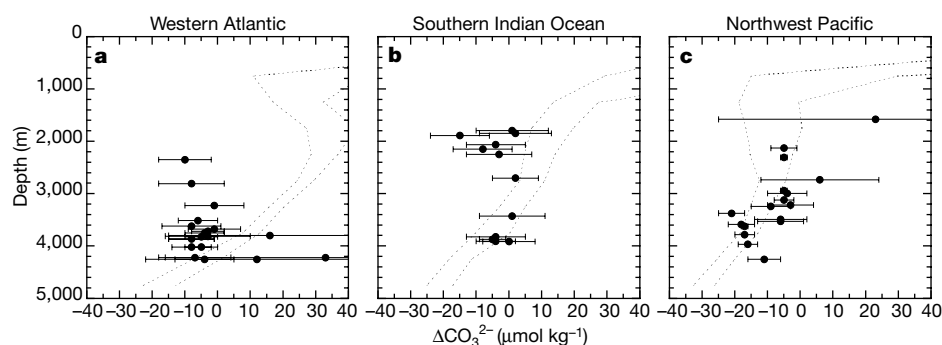


Figure 3 Vertical profiles of carbonate-ion saturation (ΔCO_3^{2-}) in three regions for the modern and Last Glacial Maximum. **a**, Western Atlantic (30°S – 10°N , 30 – 60°W). **b**, Southern Indian Ocean (60 – 20°S , 70 – 120°E). **c**, Northwest Pacific (10 – 50°N , 120°E to 150°W). The range in ΔCO_3^{2-} observed today in each region is shown by the dashed lines marking one standard deviation above and below the regional mean. Each

LGM estimate (filled circle) is derived from the ΔCO_3^{2-} associated with the five closest modern analogue samples, determined using the squared-chord dissimilarity measure. The uncertainty in LGM estimates is the standard deviation in ΔCO_3^{2-} of the five closest modern analogs.

pheric p_{CO_2} , well documented in ice cores, demands a glacial increase in CO_3^{2-} , at least at the sea surface. The low values of CO_3^{2-} that we observe at intermediate depths throughout all three basins are therefore puzzling. Perhaps the method breaks down in supersaturated intermediate waters, but there is no evidence for this in our internal tests of the modern data set (see Methods). Transport of CO_2 through the atmosphere, or dissolution of CaCO_3 , can decouple CO_3^{2-} from nutrients (and the palaeotracers listed above). Given the difficulty of explaining the mechanism behind the low p_{CO_2} of the glacial atmosphere¹⁵, we are reluctant to dismiss the apparently robust data presented here.

Regional reconstructions from the western Atlantic, the southern Indian Ocean, and the northwest Pacific are presented in Fig. 3. The changes are largest in the Western Atlantic, where ΔCO_3^{2-} decreased from 20 to zero or -10 between 2,000 and 4,000 m. The decrease in CO_3^{2-} made the Atlantic more corrosive across a broad depth range, consistent with other observations of reduced preservation relative to today. There is some indication that ΔCO_3^{2-} in deep water ($>4,000$ m) increased, but the standard deviation of the deep estimates are large. A similar pattern occurred in the southern Indian Ocean, where saturation decreased to around zero above 3,000 m relative to today, and increased slightly ($5 \mu\text{mol}$) at 4,000 m relative to present. The ΔCO_3^{2-} changes were small in the northwest Pacific. Above 3,000 m, any change is within 1σ of present conditions. Below 3,000 m, several samples indicate that ΔCO_3^{2-} was increased, and a few samples indicate little change. The glacial carbonate-ion reconstruction is consistent with the established picture of enhanced preservation in the glacial Pacific relative to today, and reduced preservation in the glacial Atlantic^{16–20}.

In general, our results appear to be inconsistent with the boron-isotope estimate that CO_3^{2-} at 3.5 km in the Pacific was $100 \mu\text{mol}$ higher than today¹; instead they imply that CO_3^{2-} changed little during LGM except in the Atlantic. It is possible that a change in the relative rain rates of organic carbon and calcite might have shifted the lysocline relative to the saturation horizon²¹, shifting also the

present relationship between dissolution intensity and ΔCO_3^{2-} that we rely on here. However, the relationship between dissolution intensity and ΔCO_3^{2-} in core-top samples is robust across broad geographic regions, and we do not suppose that such a systematic relationship could have changed so uniformly. Analogous objections to the rain ratio hypothesis have been raised based on the tight relationship between the saturation horizon and the calcium carbonate distribution (CCD) today²². The absence of a CO_3^{2-} change points to sea-floor CaCO_3 as an effective buffer of ocean pH on the ~ 10 -kyr timescales of the glacial–interglacial cycles. □

Methods

Carbonate-ion saturation

To estimate carbonate-ion saturation for a sample, we used the modern analogue technique to find the five samples in a modern database of 905 core-top samples that are most similar with regard to faunal composition as measured by the squared chord dissimilarity²³, using the 29 species common to the foraminifer equations used by CLIMAP⁵. The carbonate-ion saturation for each core-top sample was interpolated at the depth the core was taken from the gridded database produced by Archer²². We averaged the carbonate-ion saturation at the locations of the top five analogues to obtain the estimated carbonate-ion value (see example in Table 1). The principle that allows use to reconstruct both temperature and CO_3^{2-} is the observation that unique faunas are associated with discrete combinations of CO_3^{2-} and temperature. For example, a simplified case with only a few species, a fossil sample originating in a cold surface and undersaturated sea-floor environment (perhaps consisting of 100% *Neoglobobulimina pachyderma* left coiling) would find analogues (containing 100% *N. pachyderma*) from a similar environment in the core-top database (cold and undersaturated), and would be dissimilar to all core-top samples from cold-surface, saturated sea-floor environments (perhaps containing *Globigerina quinquiloba*, or other cold-water but dissolution-susceptible species in addition to *N. pachyderma*). Unlike transfer functions that attribute all the variation to one forcing variable, methods like the modern analogue technique are appropriate for relationships between (multivariate) species abundance and multivariate forcing influences.

Accuracy tests

The accuracy of the reconstructed carbonate-ion saturation can be evaluated in several ways. We compared expected and observed carbonate-ion saturation for the core-top database: $n = 905$; $r = 0.85$; standard error, $11 \mu\text{mol kg}^{-1}$; comparable to the root-mean-square (r.m.s) uncertainty between the gridded CO_3^{2-} data and the ungridded data from which it was constructed. A second test is to compare reconstructed versus observed mean ΔCO_3^{2-} profiles for the Atlantic and combined Indian and Pacific oceans. In Fig. 2, the open circles are reconstructed using faunal analogues, and the light solid line is the area-weighted average for the gridded water-column data. The agreement is everywhere within $10 \mu\text{mol kg}^{-1}$, and the analogue technique is even able to reproduce some of the observed variability in CO_3^{2-} . Third, we reconstructed ΔCO_3^{2-} for modern core-top samples not included in the calibration. The five closest analogues for one such modern core-top sample (RC12-113) are listed in Table 1. The observed $0 \mu\text{mol kg}^{-1}$ ΔCO_3^{2-} at the RC12-113 site is very close to the $-2 \mu\text{mol kg}^{-1}$ ΔCO_3^{2-} mean of the five analogues.

Measured changes in species abundance

Changes in species abundance above the saturation horizon are more subtle than are

Table 1 Modern analogues for sample RC12-113

Analog sample ID	Dissimilarity	Depth (m)	ΔCO_3^{2-} ($\mu\text{mol kg}^{-1}$)
V14-81	0.08	3,634	-1
V21-108	0.09	1,900	-2
V28-309	0.09	2,166	0
RC12-365	0.11	2,787	-4
V21-107	0.13	2,217	-4
Mean	0.10	2,540	-2

Sample obtained from 0–1 cm (core-top) in the southern Coral Sea, modern ΔCO_3^{2-} $0 \mu\text{mol kg}^{-1}$.

observed below, but even above the saturation horizon, observable variations in species abundance can be attributed to subsurface respiration-driven dissolution, which is expected to be influenced by the overlying water ΔCO_3^{2-} (ref. 24). If carbonate-ion saturation had no effect on foraminifer abundance above the lysocline, we would expect a larger standard deviation in ΔCO_3^{2-} for estimates above the lysocline because faunally similar samples should be associated with a wider range of saturation levels. No such trend in the standard deviation is observed, supporting our assumption that faunal patterns above the lysocline are also associated with narrow ($10\ \mu\text{mol kg}^{-1}$) ranges of ΔCO_3^{2-} . We note that the method seems to work in spite of regional variation in respiration-driven dissolution. Either these differences must be relatively small, or they are correlated with faunal assemblage.

Constraints on the method

In addition to the limitations of the method imposed by the uncertainty of the estimate, there are additional limitations imposed by the CLIMAP and modern analogue databases. From both the core-top and the CLIMAP LGM data sets, most of the samples are concentrated in the middle water column, with fewer at extreme high or low ΔCO_3^{2-} values. There are few core-top samples with ΔCO_3^{2-} higher than $+60\ \mu\text{mol kg}^{-1}$, so we limit our reconstruction to samples deeper than 1,000 m, where we expect the ΔCO_3^{2-} to remain below $50\ \mu\text{mol kg}^{-1}$. Most of the CLIMAP LGM samples lie between 2,000–4,000 m.

Received 3 August 2001; accepted 9 January 2002.

1. Sanyal, A., Hemming, N. G., Hanson, N. & Broecker, W. S. Evidence for a higher pH in the glacial ocean from boron isotopes in foraminifera. *Nature* **373**, 234–236 (1995).
2. CLIMAP Project Members. Seasonal reconstruction of the Earth's surface at the last glacial maximum. *Geol. Soc. Am. Map Chart Ser. MC-36*, 1–18 (1981).
3. Mix, A. C. Influence of productivity variations on long-term atmospheric CO_2 . *Nature* **337**, 541–544 (1989).
4. Ravelo, A. C., Fairbanks, R. G. & Philander, S. G. H. Reconstructing topical Atlantic hydrography using planktonic foraminifera and an ocean model. *Paleoceanography* **5**, 409–431 (1990).
5. Prell, W. L. The stability of low-latitude sea-surface temperatures: An evaluation of the CLIMAP reconstruction with emphasis on the positive SST anomalies. (Report TR025, Department of Energy, Washington DC, 1985).
6. Millero, F. J. The effect of pressure on the solubility of minerals in water and seawater. *Geochim. Cosmochim. Acta* **46**, 11–22 (1982).
7. Mucci, A. The solubility of calcite and aragonite in seawater at various salinities, temperatures, and one atmosphere total pressure. *Am. J. Sci.* **283**, 780–799 (1983).
8. Duplessy, J. C. et al. Deep water source variations during the last climatic cycle and their impact on the global deep water circulation. *Paleoceanography* **3**, 343–360 (1988).
9. Oppo, D. W. & Horowitz, M. Glacial deep water geometry: South Atlantic benthic foraminiferal Cd/Ca and $\delta^{13}\text{C}$ evidence. *Paleoceanography* **15**, 147–160 (2000).
10. Yu, E.-F., Bacon, M. P. & Francois, R. Similar rates of modern and last glacial ocean thermohaline circulation inferred from radiochemical data. *Nature* **379**, 689–694 (1996).
11. Winguth, A., Archer, D. & Maier-Reimer, E. In *Inverse Methods in Global Biogeochemical Cycles* (eds Kasibhatla, P. et al.) (AGU Press, Washington DC, 2000).
12. Kennett, J. P. & Ingram, B. L. A 20,000 year record of ocean circulation and climate change from the Santa Barbara Basin. *Nature* **377**, 510–514 (1995).
13. Ganeshram, R. S., Pedersen, T. F., Calvert, S. E. & Murray, J. W. Large changes in oceanic nutrient inventories from glacial to interglacial periods. *Nature* **376**, 755–758 (1995).
14. Altabet, M. A., Francois, R., Murray, D. M. & Prell, W. L. Climate-related variations in denitrification in the Arabian Sea from sediment $^{15}\text{N}/^{14}\text{N}$ ratios. *Nature* **373**, 506–509 (1995).
15. Archer, D., Winguth, A., Lea, D. & Mahowald, N. What caused the glacial/interglacial atmospheric pCO_2 cycles? *Rev. Geophys.* **38**, 159–189 (2000).
16. Farrell, J. W. & Prell, W. L. Climate change and CaCO_3 preservation: an 800,000 year bathymetric reconstruction from the central equatorial Pacific Ocean. *Paleoceanography* **4**, 447–466 (1989).
17. Peterson, L. C. & Prell, W. L. In *The Carbon Cycle and Atmospheric Carbon Dioxide: Natural Variations Archean to Present* (eds Sundquist, E. T. & Broecker, W. S.) 251–269 (American Geophysical Union, Washington DC, 1985).
18. Curry, W. B. & Lohmann, G. P. In *The Carbon Cycle and Atmospheric Carbon Dioxide: Natural Variations Archean to Present* (eds Sundquist, E. T. & Broecker, W. S.) 285–301 (American Geophysical Union, Washington DC, 1985).
19. Howard, W. R. & Prell, W. L. Late Quaternary CaCO_3 production and preservation in the Southern Ocean: Implications for oceanic and atmospheric carbon cycling. *Paleoceanography* **9**, 453–482 (1994).
20. Crowley, T. J. In *The Carbon Cycle and Atmospheric Carbon Dioxide: Natural Variations Archean to Present* (eds Sundquist, E. T. & Broecker, W. S.) 271–284 (American Geophysical Union, Washington DC, 1985).
21. Archer, D. & Maier-Reimer, E. Effect of deep-sea sedimentary calcite preservation on atmospheric CO_2 concentration. *Nature* **367**, 260–263 (1994).
22. Archer, D. An atlas of the distribution of calcium carbonate in deep sea sediments. *Glob. Biogeochem. Cycles* **10**, 159–174 (1996).
23. Overpeck, J. T., Webb, T. & Prentice, I. C. Quantitative interpretation of fossil pollen spectra: dissimilarity coefficients and the method of modern analogs. *Quat. Res.* **23**, 87–108 (1985).
24. Emerson, S. & Bender, M. L. Carbon fluxes at the sediment water interface of the deep sea: Calcium carbonate preservation. *J. Mar. Res.* **39**, 139–162 (1981).

Acknowledgements

We thank D. Lea, S. Lehman, R. Toggweiler and D. Sigman for helpful suggestions.

Correspondence and requests for materials should be addressed to D.M.A. (e-mail: dma@ngdc.noaa.gov).

Laser–Raman imagery of Earth's earliest fossils

J. William Schopf*, Anatoliy B. Kudryavtsev†, David G. Agresti†, Thomas J. Wdowiak† & Andrew D. Czaja*

* Department of Earth & Space Sciences, and Institute of Geophysics & Planetary Physics (Center for the Study of the Evolution and Origin of Life), University of California, Los Angeles, California 90095-1567, USA

† Astro and Solar System Physics Program, Department of Physics, University of Alabama at Birmingham, Birmingham, Alabama 35294-1170, USA

Unlike the familiar Phanerozoic history of life, evolution during the earlier and much longer Precambrian segment of geological time centred on prokaryotic microbes¹. Because such microorganisms are minute, are preserved incompletely in geological materials, and have simple morphologies that can be mimicked by nonbiological mineral microstructures, discriminating between true microbial fossils and microscopic pseudofossil 'lookalikes' can be difficult^{2,3}. Thus, valid identification of fossil microbes, which is essential to understanding the prokaryote-dominated, Precambrian 85% of life's history, can require more than traditional palaeontology that is focused on morphology. By combining optically discernible morphology with analyses of chemical composition, laser–Raman spectroscopic imagery of individual microscopic fossils provides a means by which to address this need. Here we apply this technique to exceptionally ancient fossil microbe-like objects, including the oldest such specimens reported from the geological record, and show that the results obtained substantiate the biological origin of the earliest cellular fossils known.

Over the past few decades, the rules for accepting ancient microfossil-like objects as *bona fide* Precambrian fossils have become well established. Such objects should be demonstrably biogenic, and indigenous to and syngenetic with the formation of rocks of known provenance and well-defined Precambrian age^{2,4,5}. Of these criteria, the most difficult to satisfy has proved to be biogenicity, in particular for the notably few fossil-like objects reported from Archaean (> 2,500 million years (Myr) old) deposits^{2,4}—the putative biological origin of which has been beset by controversy^{2,6–8}. This difficulty has been obviated in part by analyses of the isotopic composition of carbon in coexisting inorganic (carbonate) minerals and in whole-rock acid-resistant carbonaceous residues (kerogens), which have been used to trace the isotopic signature of biological (photoautotrophic) carbon fixation to at least ~3,500 Myr ago⁹. But because investigations of such bulk kerogen samples yield only an average value of the materials analysed, they cannot provide information about the biogenicity of individual minute objects that are claimed to be fossil.

Significant progress toward solving this problem has been made by using an ion microprobe to analyse the carbon isotopic composition of single Precambrian microfossils¹⁰. Laser–Raman spectroscopic imagery of ancient individual fossil microbes provides a means by which to extend available analytical data to a molecular level. In a previous study¹¹, we applied this technique to analyses *in situ* of Eocene fossil wood and Precambrian microscopic organic-walled fossils, and showed that it can provide insight into the molecular make-up and the fidelity of preservation of the kerogenous matter of which such fossils are composed. Because Raman spectroscopy is non-intrusive, non-destructive and particularly sensitive to the distinctive carbon signal of carbonaceous (kerogenous) organic matter, it is an ideal technique for such studies. Here we have used this technique to investigate graphitic, geochemically highly altered, dark brown to black carbonaceous filaments that have been inferred to be remnants of especially

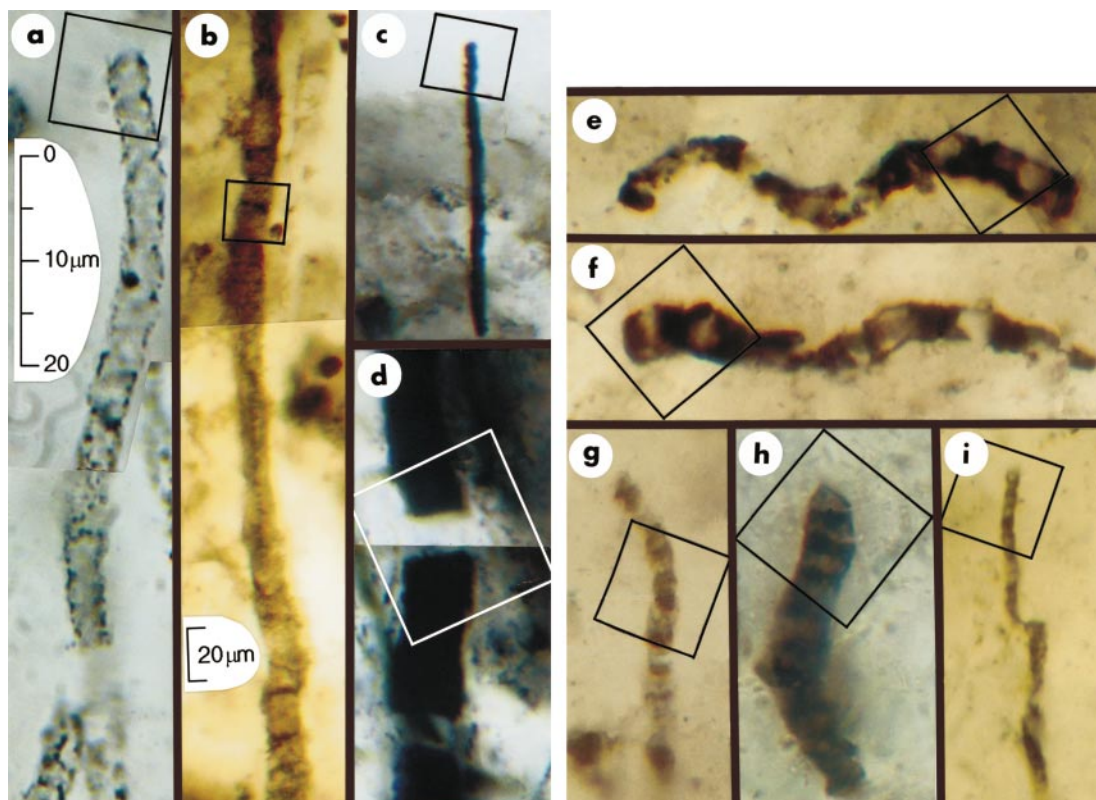


Figure 1 Optical photomicrographs showing carbonaceous (kerogenous) filamentous microbial fossils in petrographic thin sections of Precambrian cherts. Scale in **a** represents images in **a** and **c–i**; scale in **b** represents image in **b**. All parts show photomontages, which is necessitated by the three-dimensional preservation of the cylindrical sinuous permineralized microbes. Squares in each part indicate the areas for which chemical data are presented in Figs 2 and 3. **a**, An unnamed cylindrical prokaryotic filament, probably the degraded cellular trichome or tubular sheath of an oscillatoriacean cyanobacterium, from the ~770-Myr Skillogalee Dolomite of South Australia¹². **b**, *Gunflintia grandis*, a cellular probably oscillatoriacean trichome, from the ~2,100-Myr Gunflint Formation of

Ontario, Canada¹³. **c, d**, Unnamed highly carbonized filamentous prokaryotes from the ~3,375-Myr Kromberg Formation of South Africa¹⁴: the poorly preserved cylindrical trichome of a noncyanobacterial or oscillatoriacean prokaryote (**c**); the disrupted, originally cellular trichomic remnants possibly of an *Oscillatoria*- or *Lyngbya*-like cyanobacterium (**d**). **e–i**, Cellular microbial filaments from the ~3,465-Myr Apex chert of northwestern Western Australia: *Primaevifilum amoenum*^{4,5}, from the collections of The Natural History Museum (TNHM), London, specimen V.63164[6] (**e**); *P. amoenum*⁴ (**f**); the holotype of *P. delicatulum*^{4,5,15}, TNHM V.63165[2] (**g**); *P. conicoterminatum*⁵, TNHM V.63164[9] (**h**); the holotype of *Eoleptonema apex*⁵, TNHM V.63729[1] (**i**).

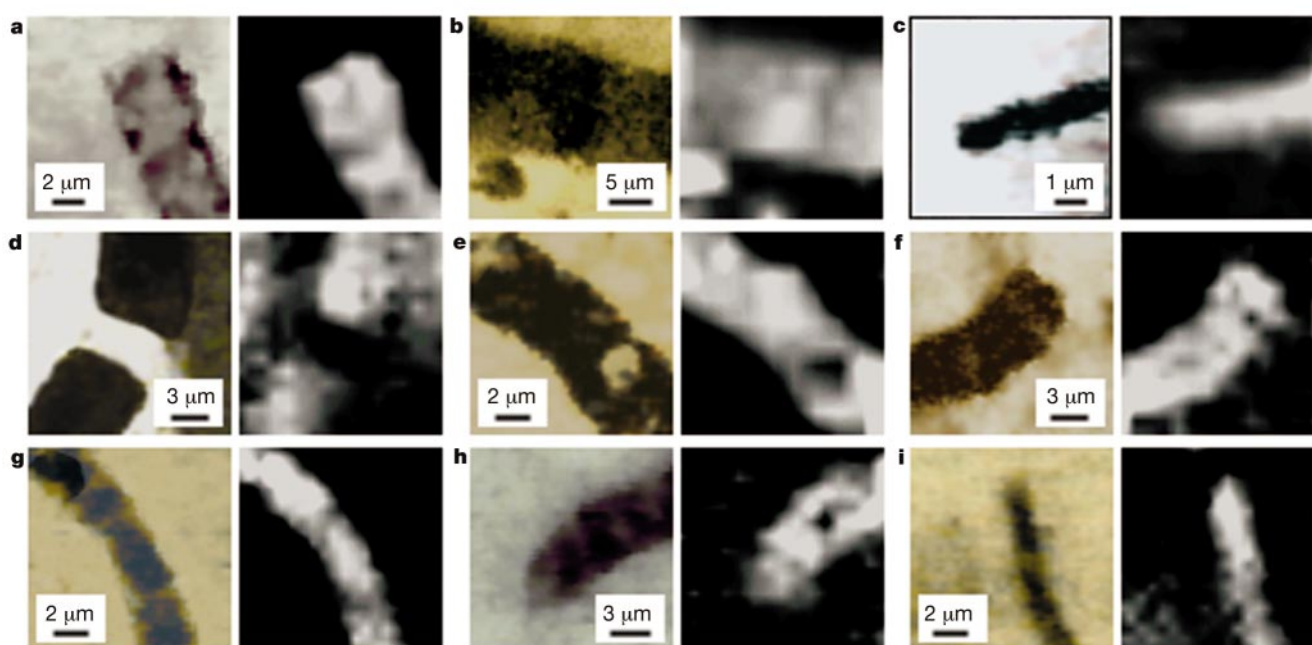


Figure 2 Digital optical images and corresponding Raman images. **a–i**, Optical images (left) and Raman 'G' band intensity maps (right), of areas of fossils indicated by the squares in Fig. 1a–i, respectively.

ancient microbes, and show that laser-Raman spectroscopic imagery can provide powerful evidence of the biogenicity of even such poorly preserved microstructures, including those regarded to be the oldest known fossils.

We analysed microbial fossil filaments that had been three-dimensionally permineralized in cryptocrystalline cherts from one subgreenschist facies (Gunflint Formation) and three geochemically more altered (greenschist facies) Precambrian geological units: (1) a cylindrical prokaryotic filament (Fig. 1a) from a domical stromatolite of the ~770-Myr Skillogalee Dolomite of South Australia¹²; (2) a cellular trichome (Fig. 1b) from a domical stromatolite of the ~2,100-Myr Gunflint Formation of Ontario, Canada¹³; (3) two prokaryotic filaments (Fig. 1c, d) from flat-laminated microbial mats of the ~3,375-Myr Kromberg Formation of South Africa¹⁴; and (4) the oldest fossils known—five specimens (Fig. 1e–i) representing 4 of 11 described taxa^{4,5} of cellular microbial filaments permineralized in organic-rich clasts of the ~3,465-Myr Apex chert of northwestern Western Australia^{4,5,15}. The range of Raman spectra obtained from fossil microbes and associated organic detritus petrified in 24 Precambrian cherts including that preserved in less geochemically altered units, and the correlation of such spectra with H/C ratios measured in bulk-sample kerogens and their relevance to processes of organic metamorphism will be described elsewhere (J.W.S. *et al.*, manuscript in preparation).

Together with optical (Fig. 1) and Raman (Fig. 2) images of fossils analysed from the four geological units, we present in Fig. 3 the dominant features of their Raman spectra: vibrational bands at ~1,350 cm⁻¹ and ~1,600 cm⁻¹, which are characteristic of carbonaceous (kerogenous) materials and commonly designated 'D' (disordered) and 'G' (graphitic), respectively, because of their presence in various forms of graphite¹⁶.

The results establish the kerogenous composition of the microscopic structures studied. Raman spectra also show that the filaments are embedded in fine-grained quartz and are devoid of virtually all other mineral phases that are identifiable by Raman analysis^{17,18}. As shown by the similarity of the spectra obtained from fossils preserved in geochemically highly altered units, for example, ~770 Myr (Fig. 3a), ~3,375 Myr (Fig. 3c, d) and ~3,465 Myr (Fig. 3e–i), the graphite-like character of the kerogen signature is related to the fidelity of preservation rather than to geological age.

The kerogen signal is not a result of contamination by the immersion oil used to enhance the optical image of some specimens, as is evident by comparing the spectra of filaments both without (Fig. 3a, b, upper panels) and with (Fig. 3a, b, lower panels) an overlying veneer of oil, as well as by the prominent kerogen signal of numerous specimens measured in the absence of such oil (for example, Fig. 3e, f, h). Indeed, not only is the Raman spectrum of the immersion oil (Fig. 3j) easily distinguishable from the kerogen signal of the fossils (for example, by the presence in the oil spectrum of major bands at ~1,000 cm⁻¹ and ~1,450 cm⁻¹) but, because the confocal capability of the analytical system allowed a vertical resolution of 1–3 µm of the signals acquired, precise focusing of the laser permitted exclusion of spectral contributions from this potential contaminant.

Similarly, the kerogen signal is not a result of contamination from graphitic (no. 2) pencil markings used in the past to encircle the Apex fossils (Fig. 1e–i) to indicate their locations. Such markings are no longer present, having been cleaned thoroughly from the upper surfaces of the Apex thin sections, and the Raman signature of pencil lead, characterized by a small broad 'D' band and a major and exceedingly sharp 'G' band¹⁷, is easily distinguishable from the signature of kerogen. In addition, rather than being exposed at upper surfaces of thin sections, the fossils studied are enclosed completely in the embedding quartz matrix and therefore have been analysed at depths (measured optically to be as much as 65 µm) in the sections where they are demonstrably not affected by surface contamination.

Finally, by correlating directly molecular composition with filament morphology, the Raman images (Fig. 2) establish unequivocally that the specimens shown are composed of carbonaceous (graphitic) kerogen and that this organic matter is present in the fossils in concentrations much greater than those in the wispy, mucilage-like clouds of finely divided particulate kerogen in which fossils of each of the four deposits are preserved.

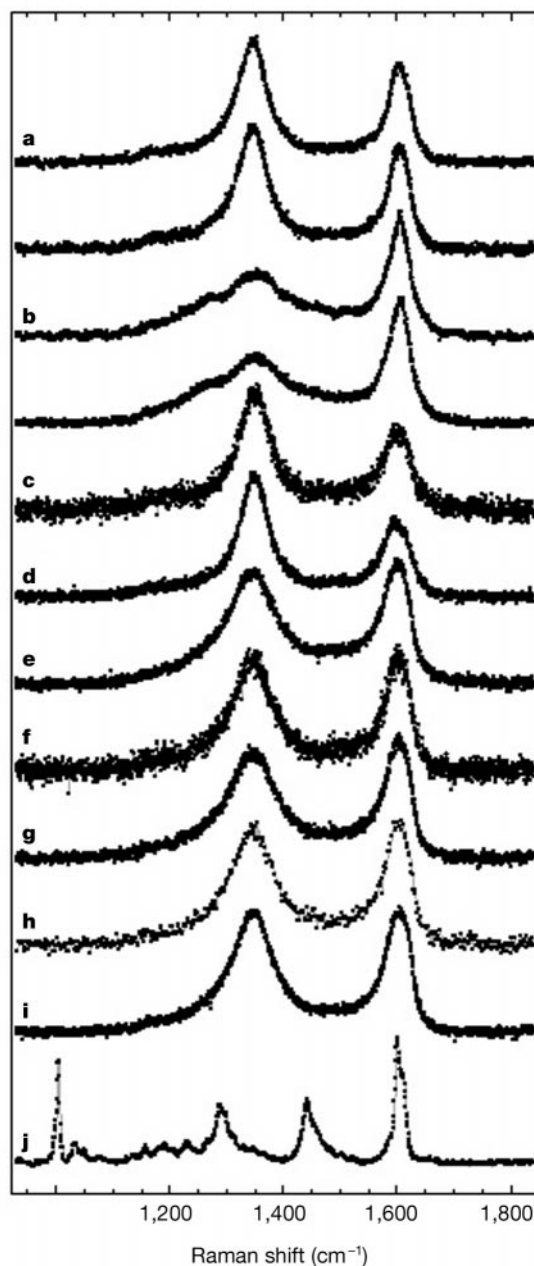


Figure 3 Typical spectral bands of parts of fossils shown in Fig. 1 used for Raman imaging (Fig. 2) and the spectrum of immersion oil used to enhance the optical image of some specimens. **a, b**, Point spectra of the Skillogalee filament shown in Fig. 1a, without (above) and with (below) a covering veneer of immersion oil; and of *Gunflintia grandis* (Fig. 1b), without (above) and with (below) an oil veneer (showing a broad 'D' band and its substructure, typical of the kerogen in this relatively unmetamorphosed, subgreenschist facies, geologic unit). **c–i**, Point spectra of other fossils from which Raman images were acquired: unnamed filamentous prokaryotes from the Kromberg Formation (**c, d**; shown in Fig. 1c, d, respectively); cellular filamentous prokaryotes from the Apex chert (**e–i**; shown in Fig. 1e–i, respectively). **j**, Spectrum of immersion oil used to enhance the optical image of some specimens.

Our data show that, applied to exceedingly ancient and geochemically highly altered (graphitic, carbonized) kerogenous microscopic fossils, Raman imagery can be used to correlate directly chemical composition with optically discernible morphology and to prove the presence of crucial indicators of biogenicity, such as kerogenous cell walls. When applied to particularly poorly preserved fossil-like microscopic filaments, for which optical microscopy can only hint at cellularity (Fig. 1a, c, d), this highly sensitive technique can yield convincing evidence of biogenicity. Most notably, by providing a means by which to correlate biologically characteristic (kerogenous) composition with features of morphology that typify filamentous microbes—such as cylindrical (Fig. 1a–i), sinuous (Fig. 1a, e, f, h) and tapering (Fig. 1e, f, h) filament form, and the presence of transverse cell walls (Fig. 1b, e–i) and distinctively shaped terminal cells (Fig. 1e, f, h)—optical microscopy (Fig. 1) coupled with laser–Raman imagery (Fig. 2) and measurement of Raman point spectra (Fig. 3) substantiates the biological origin of the oldest putative fossils now known (Fig. 1e–i). By the time of preservation of the diverse, exceptionally ancient, microscopic filaments of the Apex chert^{4,5}, ~3,500 Myr ago, microbial life was flourishing and presumably widespread. □

Methods

Optical photomicrography

We acquired the optical photomicrographs (Fig. 1) in transmitted white light with a Leitz Orthoplan 2 automatic photomicroscope.

Raman spectroscopy

Laser–Raman spectroscopic data (Figs 2 and 3) were obtained with a Dilor XY (formerly Instruments S.A., now J.Y. Horiba) 0.8-m triple-stage system that has macro-, micro- and confocal line-scan imaging options. This system permitted acquisition of both individual point spectra, typically over a period of 100 s, and true Raman images that show the spatial distribution of molecular components. The confocal capability of the system was such that we could use a 100× objective without immersion oil (having an extended working distance of 3.4 mm and a numerical aperture of 0.8) to obtain all Raman data with a planar resolution of < 1 µm and a vertical resolution of 1–3 µm. Laser wavelengths ranging from blue to infrared were provided by a coherent krypton ion laser equipped with appropriate optics. We used a wavelength of 531 nm (in the green portion of the spectrum) typically at a laser power of less than 8 mW over a 1-µm spot—an instrumental configuration that we showed to be well below the threshold that results in radiation damage to the specimens analysed.

Fossil specimens situated in the uppermost ~65 µm of petrographic thin sections were centred in the path of the laser beam projected through an Olympus BX40 microscope. The upper surfaces of thin sections were either polished (with a paste containing 1-µm-sized diamonds) or unpolished and finished by a slurry of 600 mesh carborundum. To enhance the optical image of specimens situated more than ~10 µm below the upper surface of unpolished sections, we covered the area imaged by a thin veneer (~1-µm thick) of type B non-drying microscopy immersion oil (R. P. Cargille Laboratories). Not only have appropriate analyses (see above) shown that the presence of this thin veneer has no appreciable effect on the Raman spectra acquired, but the point spectra of specimens in polished sections (for example, Fig. 1a, b) and of most specimens (12/23) analysed in unpolished sections (for examples, Fig. 1e, f, h) were obtained without this veneer.

Raman imaging

To acquire Raman images (Fig. 2), a rectangular area enclosing a part of a fossil was selected for imaging, the backscattered Raman spectra obtained in each rectangle were collected through the optical system described above along micrometre-resolution scan lines, and their x–y registrations were automatically recorded to provide a pixel-assigned array of spectral elements ('spexels')¹¹. A typical Raman image comprised 25 × 25 spectral elements, each analysed for 5–10 s, resulting in a total data collection time of 50–100 min for each image. We processed the several hundred spexels for each specimen by constructing a map of the intensity in the spectral window corresponding either to the 'D' band (at ~1,350 cm⁻¹) or the 'G' band (~1,600 cm⁻¹) of the kerogenous material; maps made by the 'G' band were generally of relatively higher Raman image quality. The resulting 'chemical image' shows the areal distribution of the fossil structures that produced the Raman spectral bands originating from specific molecular components.

Received 3 July; accepted 30 November 2001.

1. Schopf, J. W. in *The Proterozoic Biosphere, A Multidisciplinary Study* (eds Schopf, J. W. & Klein, C.) 25–39 (Cambridge Univ. Press, New York, 1992).
2. Schopf, J. W. & Walter, M. R. in *Earth's Earliest Biosphere, Its Origin and Evolution* (ed. Schopf, J. W.) 214–239 (Princeton Univ. Press, Princeton, 1983).
3. Mendelson, C. V. & Schopf, J. W. in *The Proterozoic Biosphere, A Multidisciplinary Study* (eds Schopf, J. W. & Klein, C.) 865–951 (Cambridge Univ. Press, New York, 1992).

4. Schopf, J. W. in *The Proterozoic Biosphere, A Multidisciplinary Study* (eds Schopf, J. W. & Klein, C.) 25–39 (Cambridge Univ. Press, New York, 1992).
5. Schopf, J. W. Microfossils of the Early Archean Apex chert: new evidence of the antiquity of life. *Science* **260**, 640–646 (1993).
6. Buick, R. Carbonaceous filaments from North Pole, Western Australia: are they fossil bacteria in Archean stromatolites? *Precambrian Res.* **24**, 157–172 (1984).
7. Awramik, S. M., Schopf, J. W. & Walter, M. R. Carbonaceous filaments from North Pole, Western Australia: are they fossil bacteria in Archean stromatolites? A discussion. *Precambrian Res.* **39**, 303–309 (1988).
8. Buick, R. Carbonaceous filaments from North Pole, Western Australia: are they fossil bacteria in Archean stromatolites? A reply. *Precambrian Res.* **39**, 311–317 (1988).
9. Hayes, J. M., Des Marais, D. J., Lambert, I. B., Strauss, H. & Summons, R. E. in *The Proterozoic Biosphere, A Multidisciplinary Study* (ed. Schopf, J. W. & Klein, C.) 81–134 (Cambridge Univ. Press, New York, 1992).
10. House, C. H. et al. Carbon isotopic composition of individual Precambrian microfossils. *Geology* **28**, 707–710 (2000).
11. Kudryavtsev, A. B., Schopf, J. W., Agresti, D. G. & Wdowiak, T. J. In situ laser–Raman imagery of Precambrian microscopic fossils. *Proc. Natl Acad. Sci. USA* **98**, 823–826 (2001).
12. Schopf, J. W. & Fairchild, T. R. Late Precambrian microfossils: a new stromatolitic microbiota from Boorthanna, South Australia. *Nature* **242**, 537–538 (1973).
13. Barghoorn, E. S. & Tyler, S. A. Microorganisms from the Gunflint chert. *Science* **147**, 563–577 (1965).
14. Walsh, M. M. & Lowe, D. R. Filamentous microfossils from the 3,500-Myr-old Onverwacht Group, Barberton Mountain Land, South Africa. *Nature* **314**, 530–532 (1985).
15. Schopf, J. W. & Packer, B. M. Early Archean (3.3-billion to 3.5-billion-year-old) microfossils from Warrawoona Group, Australia. *Science* **237**, 70–73 (1987).
16. Tuinstra, F. & Koenig, J. L. Raman spectra of graphite. *J. Chem. Phys.* **53**, 1126–1130 (1970).
17. Williams, K. P. J., Nelson, J. & Dyer, S. *The Renishaw Raman Database of Gemological and Mineralogical Materials*. (Renishaw Transducers Systems Division, Gloucestershire, England, 1997).
18. McMillan, P. F. & Hofmeister, A. M. Infrared and Raman spectroscopy. *Rev. Mineral.* **18**, 99–159 (1988).

Acknowledgements

We thank M. Walsh for loan of the Kromberg specimens; The Natural History Museum, London, for loan of the Apex specimens; and J. Shen-Miller for manuscript review. This work was supported by grants from the JPL/CalTech Astrobiology Center (to J.W.S.) and from the National Aeronautics and Space Administration Exobiology Program (to T.J.W.). A.D.C. is an NSF predoctoral Fellow. The Raman imaging facility at the University of Alabama at Birmingham is a consequence of the vision of L. DeLucas.

Correspondence and requests for materials should be addressed to J.W.S. (e-mail: schopf@ess.ucla.edu).

Questioning the evidence for Earth's oldest fossils

Martin D. Brasier*, Owen R. Green*, Andrew P. Jephcoat*, Annette K. Kleppe*, Martin J. Van Kranendonk†, John F. Lindsay‡, Andrew Steele§ & Nathalie V. Grassineau||

* Earth Sciences Department, University of Oxford, Parks Road, Oxford OX1 3PR, UK

† Geological Survey of Western Australia, 100 Plain Street, East Perth, Western Australia, 6004, Australia

‡ Research School of Earth Sciences, Australian National University, Canberra ACT 0200, Australia

§ School of Earth, Environmental and Physical Sciences, University of Portsmouth, Burnaby Road, Portsmouth PO1 3QL, UK

|| Department of Geology, Royal Holloway University of London, Egham Hill, Surrey TW20 0EX, UK

Structures resembling remarkably preserved bacterial and cyanobacterial microfossils from ~3,465-million-year-old Apex cherts of the Warrawoona Group in Western Australia^{1–4} currently provide the oldest morphological evidence for life on Earth and have been taken to support an early beginning for oxygen-producing photosynthesis⁵. Eleven species of filamentous prokaryote, distinguished by shape and geometry, have been put forward as meeting the criteria required of authentic Archaeal micro-

fossils^{1–5}, and contrast with other microfossils dismissed as either unreliable or unreproducible^{1,3,6,7}. These structures are nearly a billion years older than putative cyanobacterial biomarkers⁸, genomic arguments for cyanobacteria⁹, an oxygenic atmosphere¹⁰ and any comparably diverse suite of microfossils⁵. Here we report new research on the type and re-collected material, involving mapping, optical and electron microscopy, digital image analysis, micro-Raman spectroscopy and other geochemical techniques. We reinterpret the purported microfossil-like structure as secondary artefacts formed from amorphous graphite within multiple generations of metalliferous hydrothermal vein chert and volcanic glass. Although there is no support for primary biological morphology, a Fischer–Tropsch-type synthesis of carbon compounds and carbon isotopic fractionation is inferred for one of the oldest known hydrothermal systems on Earth.

The Apex microfossils were reported to occur in rounded grains of chert (microcrystalline silica) transported a great distance before redeposition in a bedded grainstone conglomerate^{2,3,5}, in a setting compared with a wave-washed beach or the mouth of a stream or river⁵. However, recent field mapping and sampling (Fig. 1) reveals

that the fossiliferous chert at Chinaman Creek is not part of the bedded succession. It is a breccia that infills one of a series of chert veins (Fig. 1, sample 4) that cross-cuts pillow basalts and feeds up into, and is continuous with, overlying stratiform cherts (Fig. 1, samples 6–8)¹¹. Textural and scanning electron microscope–energy dispersive X-ray (SEM–EDX) studies indicate a suite of hydrothermally associated minerals plus the hydrothermal alteration of adjacent pillow basalts but not overlying komatiitic basalts (Fig. 1; see Methods). A hydrothermal origin is inferred for this vein and closely associated stratiform cherts, much like that inferred for some barite–chert beds of the 3,490 million year (Myr) Dresser Formation at nearby North Pole¹².

We find no evidence for surface sedimentary structures or textures (for example, sorting or rounded quartz) in the microfossiliferous clasts. Petrography of both figured and re-collected microfossiliferous chert samples (Fig. 1, sample 4) from the chert breccia vein reveals multiple generations of fissure filling (fabric A, generations A1 to A4), brecciation, and veining by chalcedonic microquartz (fabric B1 to B3), with a final phase of void-filling by megaquartz (fabric C; Fig. 1; see Methods and Supplementary

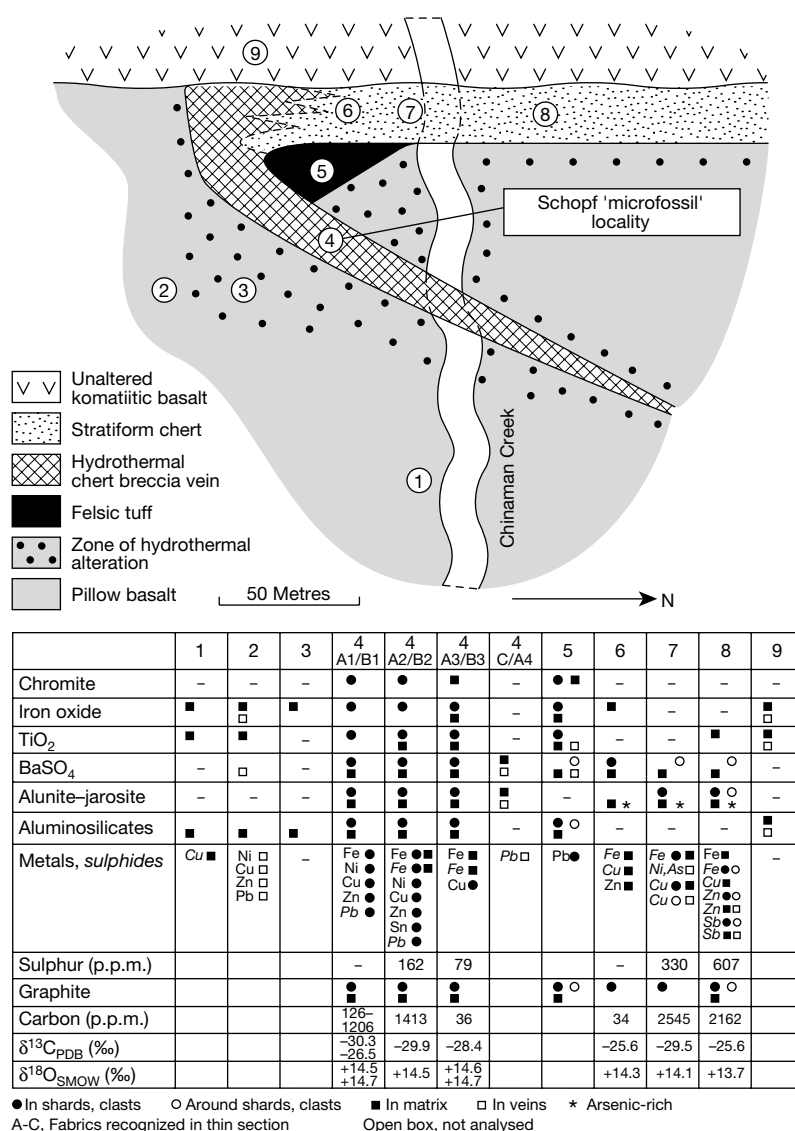


Figure 1 Geological sketch map of the Apex chert at Chinaman Creek, showing sample numbers and site of the Schopf ‘microfossil’ locality (sample 4) from a metalliferous hydrothermal chert breccia vein that cross-cuts hydrothermally altered pillow basalt.

Samples 1–9 are from this study. New geochemical data are consistent with a hydrothermal setting (see Methods).

Information Fig. 1a, b). Structures with undulose laminations (Supplementary Information Fig. 1a, S) have hitherto been regarded as stromatolite-like clasts³. However, we find that they can show intimate relationships with enclosing fabrics of A1–A3, forming void-filling drapes and overhangs. These ‘stromatoloids’ contain lithoclasts, phyllosilicates, metallic oxides, sulphides and

rhombic ghosts like those of the enclosing fabric. We reinterpret these stromatoloids as isopachous internal cements of generations B1–B3.

High-resolution stable isotopic analyses of fabrics from freshly cut rock slices reveal a range of $\delta^{13}\text{C}_{\text{PDB}}$ values (where $\delta^{13}\text{C} = [({}^{13}\text{C}/{}^{12}\text{C})_{\text{sample}}/({}^{13}\text{C}/{}^{12}\text{C})_{\text{standard}}] - 1$) of -30‰ to -26‰ for reduced carbon (see Methods and Fig. 1), which is comparable with results from earlier studies and not inconsistent with biogenic fractionation^{4,13}. Of these, the most negative $\delta^{13}\text{C}$ value is likely to be the best preserved, as indicated by dehydrogenation studies of Warrawoona kerogens elsewhere⁴. Our values of -30‰ are in the range of those obtained from other early hydrothermal cherts¹⁴. Oxygen isotopes for SiO_2 of $+13.7\text{‰}$ to $+14.7\text{‰}$ are less ^{18}O -enriched than many Archaean sedimentary cherts (compare ref. 15). Along with native metals (Fe, Ni, Cu, Zn, Sn) (Fig. 1), this suggests that comparatively high ($\sim 250\text{--}350^\circ\text{C}$) hydrothermal temperatures were reached¹⁵. The relatively high sulphur concentrations found within the stratiform cherts (330–607 p.p.m.) (Fig. 1) are shown by EDX to lie within rhombic pseudomorphs of microcrystalline quartz after sulphate or within microscopic sulphide grains that have complexly intergrown barite, alunite and jarosite (often arsenic-rich). Sulphur isotopic values of -6 to $+5\text{‰}$ $\delta^{34}\text{S}_{\text{CDT}}$ (Fig. 1) from combined sulphide–sulphate microsamples in these stratiform cherts are close to those obtained from broadly coeval sulphides and barite at the nearby North Pole, where sulphate growth resulted either from the hydrothermal alteration of evaporitic gypsum and anhydrite¹⁶ or from the reactions involving exhalative sulphur species in near-surface conditions^{11,12}.

Much of the fossil-bearing rock within the chert was reported to consist of rounded clasts and lithic fragments, of which about 5% were suggested to comprise a petrographically distinctive population yielding microfossils said to be absent from all other clasts and the surrounding matrix^{2,3}. This was interpreted to indicate that these clasts were initially preserved in an older unit, some part of which was eroded, transported and redeposited as a detrital component of the bedded chert^{2,3}. Our findings give no support to this interpretation. The figured microfossil-like structures occur in rounded to angular clasts of early fabric A1 ($\sim 40\%$), within second or later generations of fissure filling (A2, A3; $\sim 57\%$) and in later, metastable chert matrix (B3; $\sim 3\%$; see Supplementary Information Table 1), calling their primary origin into question. Comparable graphitic structures also occur in both clasts and matrix of the stratiform cherts (Fig. 2p), within glass rims of volcanic shards (Fig. 2l) and in the associated chalcedony matrix of the felsic tuffs (Fig. 2e, k), clearly calling their biogenicity into question.

The structures thought to be cellularly permineralized microfossils^{1–3} occur mainly as isolated, irregularly distributed and randomly orientated filaments in these fabrics. They therefore differ from typical populations of Proterozoic to modern cyanobacteria, in which the trichomes commonly occur clustered together in layers, often with a distinctive orientation relative to bedding (for example, ref. 17). Previous authors^{1–3} have interpreted the darker mineral of these filaments to be kerogen and the transecting paler areas to be of cellular origin, taken together to reflect a trichomic organization like that of Proterozoic to modern bacteria and cyanobacteria. Features interpreted as the terminal and medial cell shape, the cellular dimension and the degree of trichomic attenuation have then been used to distinguish 11 taxa of filamentous microfossils^{2,3}.

We have made detailed, digital automontage images of the figured material of purported microfossils (including holotypes and paratypes) and of associated, previously unillustrated filamentous microstructures (Figs 2 and 3; see Methods). A few of the filaments are sinuous and S-shaped (for example, Fig. 2c, j) but most are tightly C- or J-shaped (for example, Fig. 2m, z) or L-shaped (for example, Fig. 2r) and some are almost completely closed planar

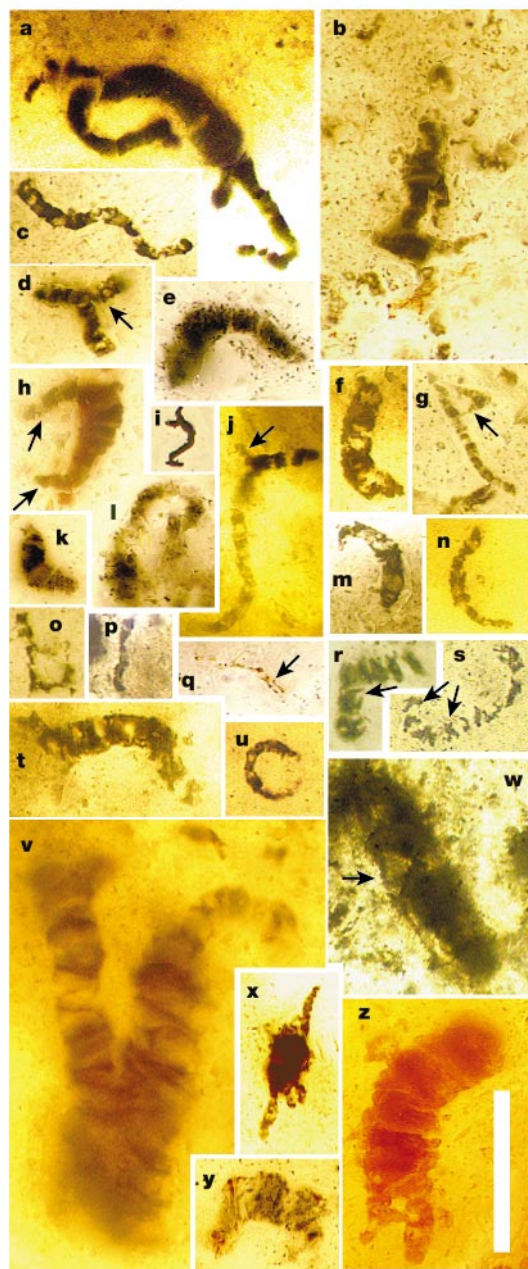


Figure 2 Automontages of inferred artefacts from the Apex chert. **a, b, o, u, v, x**, Pseudoseptate and branched filamentous artefacts from vein chert (NHM V.63127, 63164, 63165, 63127, 63164, 63164). **i**, Artificial graphite filament (63166). **e, k, l**, Filamentous artefacts from felsic tuff (**e** and **k** from chalcedony matrix; **l** from within devitrified rim of volcanic glass shard). **p**, Pseudoseptate filament from clast in stratiform chert. **c, d, f–h, j, m, n, q–t, w, y, z**, New images of previously illustrated structures (refs 2, 3; holotypes*): **c**, *Primaevifilum amoenum*; **d**, *A. disciformis*; **f**, *P. laticellulosum**; **g**, *P. delicatulum**; **h**, *P. attenuatum**; **j**, *P. amoenum**; **m**, *P. conicoterminatum**; **n**, type C narrow filament; **q**, *Archaeotrichion septatum**; **r**, **s**, ‘trichome showing bifurcated cells’; **t**, type A broad filament; **w**, cf. *A. grandis**; **y**, ‘type B’ broad filament; **z**, *Archaeosclerotioropsis maxima**. Bright-field transmitted light. Scale bar, 40 μm ; arrows indicate anomalies (see the text).

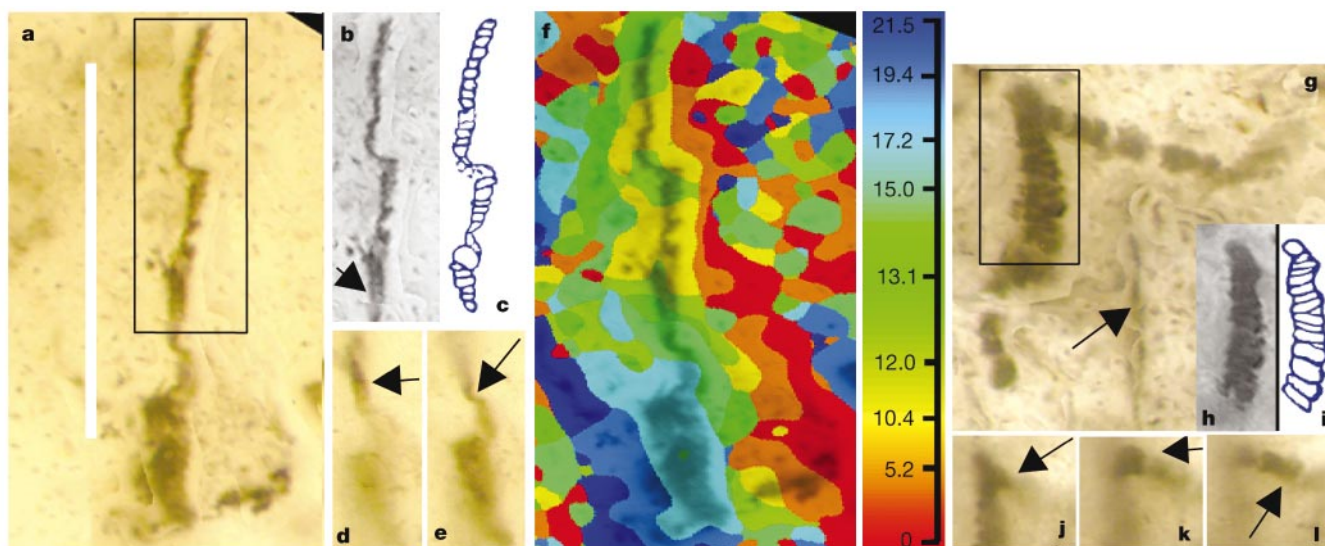


Figure 3 Automontages of inferred artefacts from the Apex chert. **a**, New image of putative beegiatoan *Eoleptonema apex* Holotype (ref. 3), combining the most sharply focused images from successive focal planes; **b, c**, digital image and interpretative sketch in the style of ref. 3 that omits the lower structure; **d, e**, new single image frames showing continuity of original and newly imaged structures. **f**, Topographic map showing

computer-selected focal planes (plus scale in μm) of **a**; **g**, new image of putative cyanobacterium *Archaeosclerolites disciformis* Holotype (ref. 3) showing rhombic ghost (arrow); **h, i**, digital image and interpretative sketch in the style of ref. 3 that omits the lower structure and side branch; **j, k, l**, new single image frames showing continuity of original and newly imaged structure. Scale bar, 40 μm .

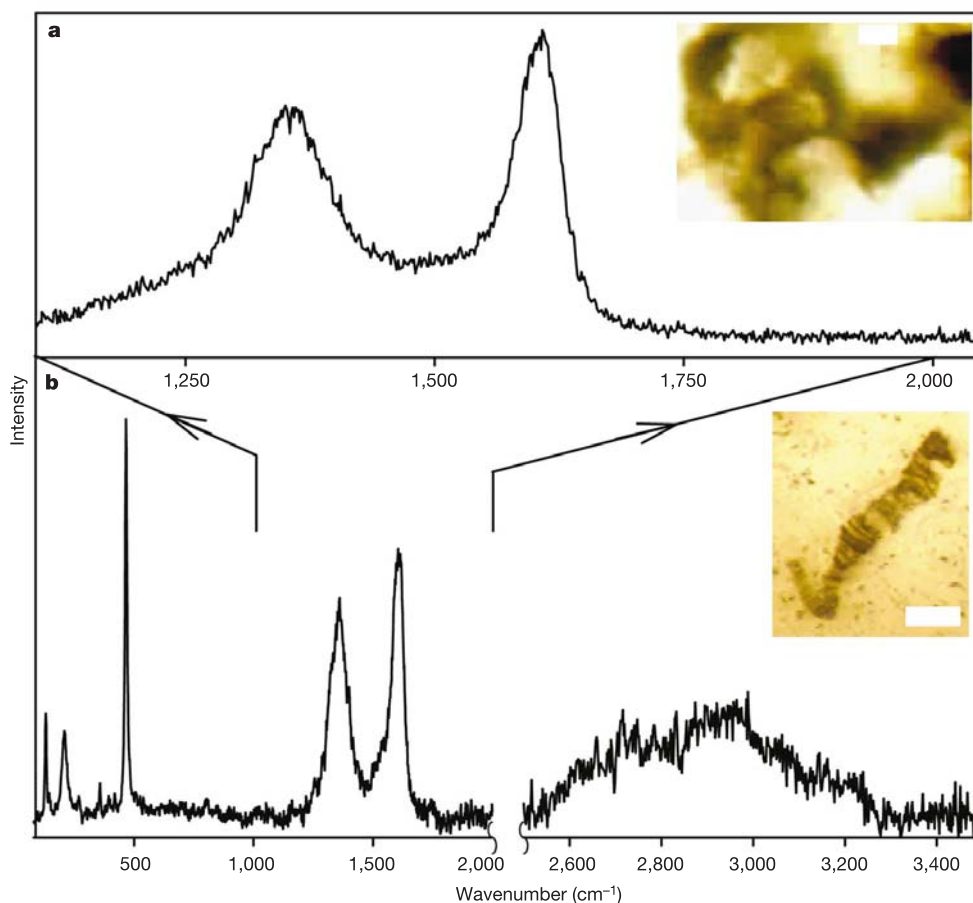


Figure 4 Raman spectra of associated graphitic objects (<1 mm apart) within NHM V.63165. **a**, Spherulitic mass in B2; **b**, pseudofossil identified as a 'degraded cellular filament or wrinkled sheath' (Fig. 1.5.4B in ref. 2) within fabric A1, showing quartz modes (128, 206, 355 and 464 cm^{-1}) plus first- and second-order graphitic carbon peaks.

Comparable spectra are also produced by dark microclasts within fabric A2 and dendritic pseudoseptate mullions within fabric B2. Background has been subtracted. Scale bar for insets, 10 μm .

circles (Fig. 2u). Many of the filamentous structures are branched or formed in ways not shown in the original descriptions and illustrations^{1–3} because of the choice of focal depth and/or illustrated field of view. In *Primaevifilum amoenum*, interpreted as a prokaryote cellular trichome (ref. 2), the holotype, found within fabric A2, is seen to have a small continuous side branch (Fig. 2j). In *Archaeosclerotriopsis disciformis* of inferred oscillatoriacean cyanobacterial affinity³, the holotype (Fig. 3g–l) is seen to be part of a continuous bifurcating structure within fabric A3 whose shape closely follows that of a nearby rhombic crystal ghost (Fig. 3g, arrow). In *P. delicatulum*, compared with modern Oscillatoriaceae², the holotype is found to be part of a complex structure involving a continuous upper side branch (Fig. 2g). Branching is inconsistent with an oscillatoriacean affinity and not generally confirmed in the fossil record until ~900–800 Myr (ref. 18). We also find a continuum between such filaments and numerous comparable but unreported pseudoseptate artefacts with multiple, chaotic branches of markedly varying diameters (Fig. 2a, b) and multiple filaments radiating from a bulbous central body (Fig. 2x) that occur nearby in the type slides.

The holotype of *Eoleptonema apex*, compared with the modern bacterium *Beggiatoa*³, is found to be part of a larger, sharply angular structure (Fig. 3a–f) that we infer to have formed around rhombic crystal ghosts within fabric A2, much as in Fig. 2o. The tiny thread-like holotype in *Archaeotrichion septatum* (Fig. 2q) is seen to be part of a larger branched structure, deflected along planes between quartz crystals, in an area of pervasive iron staining within fabric B3. One end of the holotype of *Primaevifilum attenuatum* was originally reconstructed (Fig. 5G in ref. 3) to incorporate a contiguous iron-stained fracture; the newly montaged structure is strongly arcuate (Fig. 2h) and lies within fabric A2. A further observation concerns maximum filament width, which was much greater (19.5 µm) than seen in most other Precambrian microfossil assemblages³. Our studies extend this phenomenon, revealing branched pseudoseptate structures up to 36 µm in diameter (Fig. 2v).

High-resolution Raman spectroscopy has been suggested as a technique that can clarify the carbonaceous nature of an ancient microfossil-like object of putative but uncertain biogenicity, especially if it lies too far beneath the surface for other surface analytical techniques¹⁹. Biogenicity would remain unproved, however, if associated artefacts and the groundmass were of the same composition and particle size as the filaments. That is what we find in this case. In Fig. 4 we compare high-resolution micro-Raman spectra of a figured microfossil in fabric A1 with that of an associated spherulitic artefact in fabric B2. Both show Raman modes from a matrix of microquartz (diameter 5–10 µm) plus a multiplet at 1,356 and 1,606 cm⁻¹ and weaker bands at around 2,710 and 2,935 cm⁻¹ (Fig. 4b) characteristic of first-order and second-order bands of amorphous graphite²⁰, respectively. The mode intensities increase in direct relation to the strength of apparent black coloration observed in transmitted light. Unlike concentrated samples of carbon, such as kerogen and coal, there is little direct absorption of the incident laser power, even for much higher powers than were used to collect these spectra, suggesting that the graphite is dilute and thermally protected by the bulk (quartz) host phase. The relative mode intensities^{20,21} indicate a graphite particle size of ~200 nm for groundmass, artefacts and 'filaments' alike.

Most of these filamentous structures were claimed to preserve traces of septa which, in each filament, were thought to maintain a shape and size consistent with a biological origin^{2,3}. However, micro-Raman and thin-section petrography suggest that the septate appearance of the filaments is largely created by microcrystalline quartz grains (~1–10 µm) interspersed with darker amorphous graphite that makes up the bulk of the filament (for example, Fig. 2f, g). The appearance of numerous thin septa seems to be caused by more closely packed plates of graphite and is even seen in micro-

fossil-like structures within volcanic glass (Fig. 2e, k). The impression of supposed 'bifurcated cells' or 'cell pairs' in the process of division^{2,3} (Fig. 2r, s, arrows) can be explained by irregular alternations of darker, platy graphite with paler quartz. The illusion of septation is weakest in those cases in which the filament is thicker and the quartz grains are more randomly scattered through the graphite (Fig. 2d, t). In the presumed holotype of *A. grandis*, the crudely septate appearance can be explained by rhombic crystal ghosts (for example, Fig. 2w). Chains of tiny cell-like blobs in the holotype of *Archaeotrichion septatum* (Fig. 2q) are here interpreted as small mineral grains trapped between the planar interfaces of quartz crystals.

Ancient filamentous structures should not be accepted as being of biological origin until all possibilities of their non-biological origin have been exhausted. In particular, it is important to note that complex 'septate' carbonaceous structures can result from experimental hydrothermal processes^{22,23}. Our study of the Apex chert microfossils reveals that most are part of a morphological continuum (a 'symmetry-breaking cascade') ranging from spheroidal through bulbous, radiating and chaotically branching to arcuate forms (compare refs 7 and 22). 'Holotypes' often lie close to patches of spherulitic or botryoidal (hemispherical) chalcidony with graphitic rims of identical mineralogical appearances, compositions and particle sizes in micro-Raman spectra (see Supplementary Information Table 1). Where concentrations of graphitic impurities are high, we find that secondary spherulites of silica (formed by the crystallization of glassy or colloidal silica to fibrous chalcidonic quartz) are surrounded by nearly spherical masses of graphite, producing a spongy, spheroidal fabric in thin section (for example, Fig. 4a; Supplementary Information Fig. 1c, M1). In layers where graphitic impurities are fewer, thinner dendritic to arcuate mullions of graphite separate the spherulites (compare Fig. 2a, b, d; Supplementary Information Fig. 1c, M2). In more distal regions of purer chalcidony, the localized impurities tend to be concentrated into filamentous arcuate mullions of graphite (for example, Fig. 2c–f; Supplementary Information Fig. 1c, M3). These reaction rims can have a pseudoseptate appearance owing to the contraction and microfracturing of the amorphous graphite or to the presence of small grains or radiating sheets of quartz. We attribute variations in filament size and shape to localized variations in primary texture and spherulite and botryoid size and shape. We find that similar (but non-septate) filamentous pseudofossils are mimicked by pencil graphite infilling the arcuate fractures around chalcidony spherulites and botryoids near figured specimens (Fig. 2i). Of the 11 holotypes of prokaryotic 'microfossils' defined from these rocks^{2,3}, we regard those of *E. apex* (Fig. 3a–f) and *A. septatum* (Fig. 2q) to be mineral rims that formed around crystal margins, whereas the others can be explained as arcuate, sinuous and branched mineral rims of spherulitic origin.

Reinterpretation of the Warrawoona structures as artefacts raises serious questions about the presence of cyanobacteria at 3,460 Myr, especially because the biogenic origin of many broadly coeval stromatolites²⁴ and the evolutionary context^{9,10} could be regarded as moot²⁴. Instead, we find evidence consistent with a hydrothermal setting for the Apex chert at Chinaman Creek. Carbon isotopic values from the graphitic cherts imply a significant biological contribution to the carbon cycle, perhaps from otherwise unrepresented hyperthermophile bacteria like those thought to occur as microfossils in younger Archaean rocks²⁵. That would be consistent with evidence from the sequencing of bacterial ribosomal RNA, which indicates that methanogenic archaeobacteria are arguably of much greater antiquity than cyanobacteria⁹. However, the estimated temperatures (~250–350 °C) imply a much more extreme hydrothermal habitat than currently occupied. We suggest that redox buffers (for example, native metals within the graphitic vein cherts) could have encouraged the transformation of volcanogenic CO into isotopically light carbon compounds via a Fischer–Tropsch-type

synthesis^{26,27}, which would fit well with a hydrothermal cradle for early life. □

Methods

Field mapping, petrography, SEM–EDX

Tholeiitic pillow basalts adjacent to the vein (Fig. 1, samples 2 and 3) are bleached by the effects of hydrothermal alteration and fissured by numerous small barite-rich and metalliferous chalcadonic quartz veins, like those in the main breccia vein. Pillow basalts away from the vein (Fig. 1, sample 1) as well as komatiitic basalts above the bedded chert (Fig. 1, sample 9) are less affected by hydrothermal alteration. A wedge of felsic tuff with devitrified glass shards occurs near the top of the dyke (Fig. 1, sample 5) and is interpreted as an initial volcanic eruption from the fissure now occupied by the chert breccia vein. Similar felsic tuffaceous material has been silicified within the stratiform chert (Fig. 1, sample 6).

Optical petrography was performed on re-collected material with the use of 30- μm , 150- μm and 240- μm sections. Polished slices, hydrofluoric acid (HF)-etched rock chips and HF residues were examined under Jeol-840A and Phillips XLS30S field-emission SEMs, fitted with a light-element EDX system operating at 1–15 kV. EDX analysis shows that barite (BaSO_4) is common in both vein and stratiform cherts, together with other hydrothermally associated minerals: native metals, sulphides, chalcadonic quartz and megaquartz (Fig. 1). Both vein chert and felsic tuff contain grains of chromite (FeCr_2O_4), iron oxide and TiO_2 (for example, relicts from hydrothermal leaching or fracturing of adjacent mafic extrusives) plus Al- and K-rich phyllosilicates from hydrothermally altered feldspars. The black stratiform cherts (Fig. 1, samples 6–8) comprise weakly bedded packstones or grainstones characterized by alunite-like and jarosite-like sulphates rich in Al, K, Fe and As (for example, from hydrothermal reactions involving sulphuric acid, feldspars, zeolites or sulphides). Bedding is overprinted near the vein by rhombs, now of chalcadonic quartz, containing relict grains of barite and alunite, which are indicative of oversaturated brines near a hydrothermal vent.

Fabric mapping, automontage

Generation A1 of the vein chert (Fig. 1, sample 4) comprises brecciated, angular to rounded fragments of microcrystalline quartz containing graphitic microclasts, grains of iron oxide, rutile, chromite, native metals (Fe, Ni, Cu, Zn, including beside 'holotypes'), pyrite, sericite and rhombic pseudomorphs. The latter include relict grains of barite, plus dark reaction rims of chromite and rutile (under EDX) and amorphous graphite (in micro-Raman spectra; see below). A1 clasts are commonly traversed by small cracks and vugs now of chalcadonic microquartz (fabric B1), possibly replacing massive nodular sulphate or infilling voids left by their dissolution. A1 fragments were further brecciated by successive episodes of fracturing *in situ*, so that adjacent clasts often have a 'jigsaw puzzle' fit. The cavities thus formed were lined with epitaxial rims of spherulitic and fan-shaped white chalcadonic microquartz (fabric B2). Fissure-filling chert of generation A2 (paler with more distinct lithoclasts) was produced by fracturing and spalling of A1 and B1, geotectonically infilling voids within the breccia (Supplementary Information Fig. 1a). A comparable process led to generations A3 and B3 (Supplementary Information Fig. 1b). Conspicuous rhombic crystals (barite in EDX) or voids float in the outer margins of cross-cutting B3, which penetrates clasts of A1 and A2 via small fissures and vugs, forming laminated layers (Supplementary Information Fig. 1a) and botryoids. Each phase was accompanied by carbonaceous (hydrocarbon?) migration, penetrating along microfissures to line or infill small vugs within A1, coat and darken clasts of A1–A3 and B1–B2, and accumulate within botryoidal–spherulitic cherts of B2–B3. The central zone of wider chalcadonic veins was later infilled by crystals of drusy quartz, largely lacking in graphite (Supplementary Information Fig. 1a, b: C). Late-stage metallic oxides, often associated with secondary oxidation and reddening, infiltrate small cracks and fissures. Neither Raman nor SEM–EDX studies revealed the presence of carbonate.

Rock slices were examined under bright-field transmitted and reflected light with Nikon Optiophot-2 (biological) and Optiophot-pol (polarizing) microscopes, imaged with a single-chip RGB camera and processed with Acquis-Pro and Automontage image-capturing software. Automontage captures a succession of still digital images of three-dimensional structures at successive focal planes (Fig. 3), combining the most-sharply focused images into a single montage (Fig. 3a, g), showing the planes used (Fig. 3f).

Micro-Raman spectroscopy and stable isotopes

Micro-Raman spectroscopy, with spatial-filtering confocal aperture, was performed on rock slices, polished thin sections and type slides. Raman spectra were recorded in both 135° scattering and 180° backscattering geometry with a SPEX Triplemate equipped with a back-illuminated, liquid- N_2 -cooled charge-coupled-device detector. Spectra were excited by the 488.1-nm and 514.5-nm lines of an Ar^+ laser, focused down to a 1–5- μm spot on the sample, with powers of ~ 10 –30 mW. This power range is well below the thermal-damage threshold from local heating by the laser beam and below the threshold for any laser-induced downshift in the measured vibrational frequencies²⁸. Carbon and sulphur isotopes and concentrations (mostly replicate) were obtained from coarsely crushed, hand-picked samples of re-collected material and analysed with a He continuous-flow isotope ratio mass spectrometer²⁹ consisting of an elemental analyser Carbo Erba EA 1500 on line to an Optima mass spectrometer. Oxygen isotopes were obtained with a Synrad CO_2 laser– BrF_3 fluorination automatic system on line to an Optima mass spectrometer (method adapted from ref. 30).

Received 14 February 2001; accepted 24 January 2002.

- Schopf, J. W. & Packer, B. M. Early Archean (3.3 billion to 3.5 billion-year-old) microfossils from Warrawoona Group, Australia. *Science* **237**, 70–73 (1987).
- Schopf, J. W. in *The Proterozoic Biosphere: a Multidisciplinary Study* (eds Schopf, J. W. & Klein, C.) 25–39 (Cambridge University Press, Cambridge, 1992).
- Schopf, J. W. Microfossils of the Early Archean Apex Chert: new evidence of the antiquity of life. *Science* **260**, 640–646 (1993).
- Schopf, J. W. in *Early Life on Earth* (ed. Bengtson, S.) 193–206 (Columbia University Press, New York, 1994).
- Schopf, J. W. *The Cradle of Life* (Princeton Univ. Press, New York, 1999).
- Buick, R., Dunlop, J. S. R. & Groves, D. I. Stromatolite recognition in ancient rocks: an appraisal of irregularly laminated structures in an early Archean chert-barite unit from North Pole, Western Australia. *Alcheringa* **5**, 161–181 (1981).
- Buick, R. Microfossil recognition in Archean rocks: an appraisal of spheroids and filaments from a 3500 M.Y. old chert-barite unit at North Pole, Western Australia. *Palaios* **5**, 441–459 (1990).
- Summons, R. E., Jahnke, L. L., Hope, M. & Logan, G. A. 2-Methylhopanoids as biomarkers for cyanobacterial oxygenic photosynthesis. *Nature* **400**, 554–557 (1999).
- Hedges, S. B. et al. A genomic timescale for the origin of eukaryotes. *BioMed Central Evol. Biol.* **1**, article 4, 1–10 (2001).
- Catling, D., Zahnle, K. J. & McKay, C. P. 2001. Biogenic methane, hydrogen escape, and the irreversible oxidation of early Earth. *Science* **293**, 839–843 (2001).
- Van Kranendonk, M. J. Volcanic degassing, hydrothermal circulation and the flourishing of life on Earth: new evidence from the c. 3.45 Ga Warrawoona Group, Pilbara Craton, Western Australia. *Precambrian Res.* (in press).
- Nijman, W., De Bruin, K. & Valkering, M. Growth fault control of early Archean cherts, barite mounds, and chert-barite veins, North Pole Dome, Eastern Pilbara, Western Australia. *Precamb. Res.* **88**, 25–52 (1998).
- Strauss, H. & Moore, T. B. in *The Proterozoic Biosphere: a Multidisciplinary Study* (eds Schopf, J. W. & Klein, C.) 711–798 (Cambridge University Press, Cambridge, 1992).
- de Ronde, C. E. J. & Ebbersen, T. W. 3.2 b.y. of organic compound formation near sea-floor hot springs. *Geology* **24**, 791–794 (1996).
- Robert, F. Carbon and oxygen isotope variations in Precambrian cherts. *Geochim. Cosmochim. Acta* **52**, 1473–1478 (1988).
- Shen, Y., Buick, R. & Canfield, D. E. Isotopic evidence for microbial sulphate reduction in the early Archean era. *Nature* **410**, 77–81 (2001).
- Awramik, S. M. & Semikhatov, M. A. The relationship between morphology, microstructure, and microbiota in three vertically intergrading stromatolites from the Gunflint Iron Formation. *Can. J. Earth Sci.* **16**, 484–495 (1979).
- Mendelson, C. V. & Schopf, J. W. in *The Proterozoic Biosphere: a Multidisciplinary Study* (eds Schopf, J. W. & Klein, C.) 867–951 (Cambridge University Press, Cambridge, 1992).
- Kudryatsev, A. B., Schopf, J. W., Agresti, D. G. & Wdowiak, T. J. In situ laser-Raman imagery of Precambrian microscopic fossils. *Proc. N. Am. Acad. Sci.* **98**, 823–826 (2001).
- Wopenka, B. & Pasteris, J. D. Structural characterization of kerogens to granulite-facies graphite: applicability of Raman microprobe spectroscopy. *Am. Mineralogist* **78**, 533–557 (1993).
- Tuinstra, F. & Koenig, J. L. Raman spectrum of graphite. *J. Chem. Phys.* **53**, 1126–1130 (1970).
- Oehler, J. H. Hydrothermal crystallization of silica gel. *Bull. Geol. Soc. Am.* **87**, 1143–1152 (1976).
- Baker, R. T. K. & Harris, P. in *Chemistry and Physics of Carbon* (ed. Walker, P. L. & Throver, P. A.) 2–165 (Dekker, New York, 1978).
- Grotzinger, J. P. & Rothman, D. H. An abiotic model for stromatolite morphogenesis. *Nature* **383**, 423–425 (1996).
- Westall, F. et al. Early Archean fossil bacteria and biofilms in hydrothermally-influenced sediments from the Barberton greenstone belt, South Africa. *Precamb. Res.* **106**, 93–116 (2001).
- Holm, N. G. & Charlou, J. L. Initial indicators of abiotic formation of hydrocarbons in the Rainbow ultramafic hydrothermal system, Mid-Atlantic Ridge. *Earth Planet. Sci. Lett.* **191**, 1–8 (2001).
- Lancet, M. S. & Anders, E. Carbon isotope fractionation in Fischer–Tropsch synthesis and in meteorites. *Science* **170**, 980–982 (1970).
- Kagi, H. et al. Proper understanding of down-shifted Raman spectra of natural graphite: direct estimation of laser-induced rise in sample temperature. *Geochim. Cosmochim. Acta* **58**, 3527–3530 (1994).
- Matthews, D. E. & Hayes, J. M. Isotope-ratio-monitoring gas chromatography – mass spectrometry. *Anal. Chem.* **50**, 1465–1473 (1978).
- Clayton, R. N. & Mayeda, T. K. The use of bromine pentafluoride in the extraction of oxygen from oxides and silicates for isotopic analysis. *Geochim. Cosmochim. Acta* **27**, 43–52 (1963).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com>).

Acknowledgements

We thank C. A. Stoakes, A. T. Brasier and D. Huston for assistance with field work; N. Charnley, D. Sansom and A. T. Brasier for laboratory support; the Natural History Museum, London, for the loan of the type slides and re-collected material; R. Buick, J. Farmer, J. P. Grotzinger, A. H. Knoll, E. Nisbet, S. Moorbath, J. W. Schopf and R. E. Summons for comments on earlier versions of the manuscript; and The Royal Society, NASA Astrobiology Institute and The Carnegie Institution of Washington for support. This paper is published by permission of the Director of the Geological Survey of Western Australia.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.D.B. (e-mail: martinb@earth.ox.ac.uk).

Arctic microorganisms respond more to elevated UV-B radiation than CO₂

David Johnson*, Colin D. Campbell†, John A. Lee*, Terry V. Callaghan*‡ & Dylan Gwynn-Jones§

* Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, UK

† The Macaulay Land Use Research Institute, Craigiebuckler, Aberdeen AB15 8QH, UK

‡ Abisko Scientific Research Station, SE 981-07, Abisko, Sweden

§ Institute of Biological Science, University of Wales, Aberystwyth SY23 3DA, UK

Surface ultraviolet-B radiation and atmospheric CO₂ concentrations have increased as a result of ozone depletion and burning of fossil fuels^{1,2}. The effects are likely to be most apparent in polar regions³ where ozone holes have developed and ecosystems are particularly sensitive to disturbance⁴. Polar plant communities are dependent on nutrient cycling by soil microorganisms, which represent a significant and highly labile portion of soil carbon (C) and nitrogen (N). It was thought⁵ that the soil microbial biomass was unlikely to be affected by exposure of their associated plant communities to increased UV-B. In contrast, increasing atmospheric CO₂ concentrations were thought to have a strong effect as a result of greater below-ground C allocation⁶. In addition, there is a growing belief that ozone depletion is of only minor environmental concern because the impacts of UV-B radiation on plant communities are often very subtle⁷. Here we show that 5 years of exposure of a subarctic heath to enhanced UV-B radiation both alone and in combination with elevated CO₂ resulted in significant changes in the C:N ratio and in the bacterial community structure of the soil microbial biomass.

We used an established experiment situated close to the Abisko Scientific Research Station, Swedish Lapland (68.35°N, 18.82°E, 360 m above sea level) in a subarctic heath. The site has an open canopy of *Betula pubescens* ssp. *czerepanovii* and a dense dwarf shrub layer with scattered herbs and grasses. Plots were exposed to factorial combinations of UV-B radiation (simulating 15% ozone depletion under clear sky conditions) and elevated CO₂ (600 ± 50 p.p.m.). We determined soil microbial biomass C (C_{mic}) and N (N_{mic}) using the fumigation-extraction procedure^{8,9} and the patterns of C source utilization using extractable bacteria in

customized Biolog plates containing a range of ecologically relevant C sources¹⁰.

C_{mic} decreased significantly from 3.1 mg C per g soil dry weight in the control to 1.5 mg C per g soil dry weight in the UV-B treatment (Table 1) but remained unaffected by the application of elevated CO₂ in combination with increased UV-B. This apparent ameliorative effect may reflect differences in the quality and quantity of C entering the soil in the UV-B only and UV-B with CO₂ treatments. In contrast, N_{mic} increased from approximately 0.1 mg N per g soil dry weight in the control to 0.3 mg N per g soil dry weight in the CO₂ + UV-B treatment (Table 1). The overall main effect of the UV-B treatment was to significantly increase N_{mic} by over 100% from 0.12 to 0.27 mg N per g soil dry weight (Table 1). These changes were reflected in the microbial biomass C:N ratio ($C_{mic}:N_{mic}$) which decreased significantly by 320% in the plots receiving enhanced UV-B, but was not influenced by exposure to elevated CO₂ (Table 1).

The average well colour development (AWCD; see Methods) of all C sources in the Biolog plates was significantly lower ($P < 0.05$) in all of the treatments compared to the controls (Table 1), suggesting either a lower or less active population of bacteria. Multivariate analysis of the sole C source utilization tests showed significant discrimination ($P = 0.019$) between the treatments (Fig. 1) indicating that the soil microbial community structure had also been affected by the treatments. The plots receiving either ambient or elevated CO₂, regardless of UV-B, tended to be separated primarily by the first canonical variate, while the combined addition of increased UV-B, resulted in separation owing to both the first and second canonical variates (Fig. 1). Thus, the combination of elevated CO₂ and UV-B resulted in a population structure different to that from the individual treatments. These substantial shifts in patterns of C utilization among the four treatments will almost certainly have resulted from a change in the dominant bacterial species extracted from this soil. These observations are independent of changes in the abundance of microorganisms, as reflected by AWCD and C_{mic} (Table 1), and further demonstrate the sensitivity of the microbial component of the ecosystem.

The results from our experiments demonstrate that, unlike the plant community, the soil microbial biomass is highly sensitive to elevated UV-B radiation and CO₂ concentrations. More importantly, the impacts of the UV-B treatment on the accumulation of N in the microbial biomass may have far-reaching implications for the supply of N to plants, because the productivity of many semi-natural ecosystems is limited by N (ref. 11). We are uncertain whether these changes reflect either increased microbial

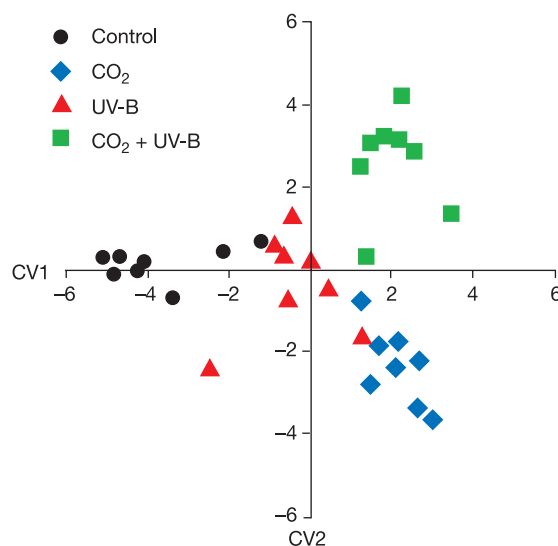


Figure 1 Ordination plot of canonical variates 1 and 2 (CV1, CV2) for sole carbon source data obtained from customized Biolog MT plates.

Table 1 Microbial responses to elevated UV-B and CO₂

Treatment	<i>C_{mic}</i>	<i>N_{mic}</i>	<i>C_{mic}:N_{mic}</i>	AWCD
Control	2.7 ^a ± 0.38	0.13 ^a ± 0.26	35.7 ^a ± 10.5	0.20 ^a ± 0.07
CO ₂	3.1 ^a ± 0.56	0.10 ^a ± 0.28	36.7 ^a ± 7.6	0.12 ^b ± 0.06
UV-B	1.0 ^b ± 0.24	0.22 ^b ± 0.42	4.8 ^a ± 1.4	0.15 ^b ± 0.06
CO ₂ + UV-B	3.7 ^a ± 0.32	0.31 ^a ± 0.21	11.3 ^a ± 1.6	0.14 ^b ± 0.05
Main effect of ambient and increased UV-B				
Ambient UV-B	2.9 ^x ± 0.47	0.12 ^x ± 0.27	36.2 ^x ± 9.1	0.16 ^x ± 0.07
Increased UV-B	2.4 ^x ± 0.28	0.27 ^y ± 0.32	8.2 ^y ± 1.5	0.14 ^x ± 0.06

Microbial biomass (mg per g soil dry weight) of carbon (*C_{mic}*) and nitrogen (*N_{mic}*), *C_{mic}:N_{mic}* ratio and average well colour development (AWCD; *A₅₉₀*) for combined carbon sources in customized Biolog plates (mean ± s.e.). Values sharing a letter within columns are not significantly different (*P* > 0.05). Individual treatments and main effects are compared separately.

accumulation of N or a change in the species composition of the microbial biomass to one with an overall lower C:N ratio. The shifts in C source utilization (Fig. 1) lend support to the second hypothesis. The capacity for subarctic semi-natural heaths to act as major sinks for fossil fuel-derived carbon dioxide is likely to be critically dependent on the supply of N, as recently reported in loblolly pine stands^{12,13}. Exposure of arctic plant communities to long-term UV-B radiation may thus alter the amount of N held within the soil microbial biomass. A full understanding of the impacts of global climate change, and in particular increased UV-B radiation, on polar plant communities can only be achieved by increasing our studies of the key soil organisms and processes that drive these ecosystems. □

Methods

Site description

The vegetation contains a layer of mosses (such as *Hylocomium splendens* (Hedw.) Br. Eur.) and lichens (such as *Peltigera aphthosa* (L.) Willd.). The major dwarf shrub species are *Empetrum hermaphroditum* Hagerup, *Vaccinium myrtillus* L., *Vaccinium uliginosum* L. and *Vaccinium vitis-idaea* L. (ref. 14). The soil is a nutrient-poor podzol on well-drained acid moraine and has a pH (H₂O) of 3.4 to 5.6 (ref. 15).

Experimental design

We erected 16 (eight controls and eight simulating enhanced UV-B radiation) metal frames (2.5 × 1.3 × 1.5 m high) each holding six UV-B fluorescent tubes during the spring of 1993. Fluorescent tubes were used to simulate 15% ozone depletion under clear sky conditions. This represents a 25% increase above ambient levels. The actual ozone depletion under typical cloud conditions at Abisko in 1994 has been calculated to be 19% (ref. 16). Timers controlled the UV-B lamps in a stepwise 'square wave' manner to account for midday increases in photosynthetically active radiation (PAR) and ambient UV-B (maximum 5.5 kJ day⁻¹)¹⁶. The daily exposure time was changed every two weeks to follow the seasonal levels of UV-B radiation at Abisko. Before use, the lamps were pre-burnt to ensure a stable UV-B output. The lamps in the control frames had glass panels to filter out UV-B and UV-C radiation. The enhanced UV-B radiation plots had ultraviolet-transmitting Plexiglas (Röhm 2458, Röhm GmbH) and cellulose diacetate (Courtaulds) to remove UV-C radiation. This was regularly replaced throughout the growing season. All plots received ambient levels of UV-A and UV-B radiation. The mean daily air temperature inside the chambers was typically 1 °C higher than that outside¹⁷.

Treatment application

Elevated CO₂ was introduced into the plots using open top chambers (0.5 m high, 0.73 m² area) constructed from Plexiglas (Röhm 2458, Röhm GmbH) placed beneath each of the frames. Air was supplied to the base of each chamber from an independent fan system (Soler and Palau). Elevated CO₂ was achieved by bleeding pure CO₂ into the air supply of half of eight of the chambers in order to maintain concentrations at 600 ± 50 μl l⁻¹. Concentrations of CO₂ within the chambers were monitored using an infrared gas analyser (Series 2000, ADC Bioscientific)¹⁷. Treatments were applied each year (from 1993 onwards) during the snow-free months of the subarctic summer from late May to early September.

Soil microbial biomass carbon and nitrogen

Duplicate cores (4 cm diameter, 6 cm deep) were extracted from each plot, the surface litter removed and the bulk soil sieved (2 mm). Microbial biomass C (*C_{mic}*) and N (*N_{mic}*) were determined in 5-g subsamples using the fumigation-extraction procedure^{8,9}. Total organic C (TOC) in the fumigated (24-h exposure to ethanol-free CHCl₃, previously washed six times with distilled water) and unfumigated extracts (0.5M K₂SO₄) was determined by sodium persulphate/ultraviolet absorption (LabTOC, Pollution and Process Monitoring). For *N_{mic}*, 2-ml subsamples of the extracts were acid-digested¹⁸ and the total N determined¹⁹. Both *C_{mic}* and *N_{mic}* were calculated using the formula: *C_{mic}* or

N_{mic} = (*X_f* - *X_{uf}*)/*K_c*, where *X_f* and *X_{uf}* are the C or N concentrations in fumigated and unfumigated extracts, respectively, and *K_c* is the proportion of microbial C (0.38; ref. 8) or N (0.54; ref. 9) extracted from soil. The *C_{mic}* and *N_{mic}* data sets were combined to produce a microbial biomass C:N ratio.

Sole carbon source utilization

Direct incubation of soil suspensions in Biolog microtitre plates containing different carbon sources in individual wells was used to determine (1) changes in the potential rate of C-source utilization, and (2) changes in relative and absolute rates of utilization of individual substrates to discriminate between soil microbial communities^{10,20}. We tested 30 ecologically relevant C sources using customized Biolog MT type plates in which the wells contained six amino acids, two carbohydrates, eight carboxylic acids, nine phenolic acids and four long-chain aliphatic acids¹⁰.

We analysed data using two different approaches. In the first, absolute rates of colour development measured as absorbance at 590 nm (*A₅₉₀*) were compared for individual C sources. The average well colour development (AWCD) during the initial 96 h of incubation was compared using two-way analysis of variance (ANOVA) and least significant difference (LSD) multiple comparison tests. In the second, multivariate analysis of the *A₅₉₀* values at equivalent AWCD from different times of incubation were compared. They were also transformed by dividing by the AWCD to avoid bias between samples with different inoculum density¹⁰. The absorbance data were analysed by canonical variate analysis, after first reducing the dimensionality by principal component analysis and by comparison of mean intergroup Mahalanobis distances with simulated confidence limits and Monte-Carlo testing of significance¹⁰. Simulated confidence limits for four groups (treatments) with eight replicates were 2.26 and 2.57 at the 95% and 99% confidence limits, respectively. All ANOVA, regression and multivariate analyses were conducted using Genstat 5.4 (NAG).

Received 7 September; accepted 6 December 2001.

- Müller, R. *et al.* Severe chemical ozone loss in the Arctic during the winter of 1995–96. *Nature* **398**, 709–712 (1997).
- Houghton, J. T. *et al.* (eds) *Climate Change 1995: The Science of Climate Change* (Cambridge Univ. Press, Cambridge, 1996).
- Caldwell, M. M., Robberecht, R. & Nowak, R. S. Differential photosynthetic inhibition by ultraviolet-B radiation in species from the arctic-alpine lifezone. *Arct. Alp. Res.* **14**, 195–202 (1982).
- Callaghan, T. V. & Jonasson, S. Arctic terrestrial ecosystems and environmental change. *Phil. Trans. R. Soc. Lond. A* **352**, 259–276 (1995).
- Moorhead, D. L. & Callaghan, T. V. Effects of increasing ultraviolet-B radiation on decomposition and soil organic matter dynamics: a synthesis and modelling study. *Biol. Fert. Soils* **19**, 19–26 (1994).
- Hungate, B. A. *et al.* The fate of carbon in grasslands under carbon dioxide enrichment. *Nature* **388**, 576–579 (1997).
- Caldwell, M. M., Searles, P. S., Flint, S. D. & Barnes, P. W. in *Physiological Plant Ecology* (eds Press, M. C., Scholes, J. D. & Barker, M. G.) 241–262 (Blackwell, Oxford, 1999).
- Vance, E. D., Brookes, P. C. & Jenkinson, D. S. An extraction method for measuring microbial biomass-C. *Soil Biol. Biochem.* **19**, 703–707 (1987).
- Brookes, P. C., Landman, A., Pruden, G. & Jenkinson, D. S. Chloroform fumigation and the release of soil-nitrogen—a rapid direct extraction method to measure microbial biomass nitrogen in soil. *Soil Biol. Biochem.* **17**, 837–842 (1985).
- Campbell, C. D., Grayston, S. J. & Hirst, D. J. Use of rhizosphere carbon sources in sole carbon source tests to discriminate soil microbial communities. *J. Microbiol. Methods* **30**, 33–41 (1997).
- Ellenberg, H. *Vegetation Ecology of Central Europe*. 4th edn (Cambridge Univ. Press, Cambridge, 1988).
- Schlesinger, W. H. & Lichter, J. Limited storage in soil and litter of experimental forest plots under increased atmospheric CO₂. *Nature* **411**, 466–469 (2001).
- Oren, R. *et al.* Soil fertility limits carbon sequestration by forest ecosystems in a CO₂-enriched atmosphere. *Nature* **411**, 469–472 (2001).
- Johanson, U., Gehrke, C., Björn, L. O., Callaghan, T. V. & Sonesson, M. The effects of enhanced UV-B radiation on a subarctic heath ecosystem. *Ambio* **24**, 106–111 (1995).
- Sonesson, M. & Lundberg, B. Late quaternary forest development of the Torneträsk area, North Sweden. 1. Structure of modern forest ecosystems. *Oikos* **25**, 121–133 (1974).
- Björn, L. O. & Holmgren, B. Monitoring and modelling of the radiation climate at Abisko. *Ecol. Bull.* **45**, 204–209 (1996).
- Gwynn-Jones, D., Lee, J. A. & Callaghan, T. V. Effects of enhanced UV-B radiation and elevated carbon dioxide concentrations on a sub-arctic forest heath ecosystem. *Plant Ecol.* **128**, 242–249 (1997).
- Bremner, J. M. & Mulvaney, C. S. in *Agronomy 9: Methods of Soil Analyses. Part 2. Chemical and Microbiological Properties* (eds Page, A. C., Miller, R. H. & Keeney, D. R.) 595–624 (Soil Science Soc. America, Madison, 1982).
- Scheiner, D. Determination of ammonia and Kjeldahl nitrogen by indophenol method. *Water Res.* **10**, 31–36 (1976).
- Garland, J. L. & Mills, A. L. Classification and characterisation of heterotrophic microbial communities on the basis of patterns of community-level sole carbon-source utilisation. *Appl. Environ. Microbiol.* **57**, 2351–2359 (1991).

Acknowledgements

We thank D. Mackinnon for TOC analysis and S. J. Chapman, L. A. Gilbert and J. R. Leake for their comments. This work was supported by the Natural Environment Research Council, the European Union and the Swedish Academy of Sciences. C.D.C. is funded by the Scottish Executive, Environment and Rural Affairs Department.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.J. (e-mail; D.Johnson@Shef.ac.uk).

Diversity-dependent production can decrease the stability of ecosystem functioning

Andrea B. Pfisterer & Bernhard Schmid

Institut für Umweltwissenschaften, Universität Zürich, Winterthurerstr. 190, CH-8057 Zürich, Switzerland

There is concern that species loss may adversely affect ecosystem functioning and stability. But although there is evidence that biodiversity loss can lead to reductions in biomass production^{1–4}, there is no direct evidence that biodiversity loss affects ecosystem resistance (ability to withstand perturbation) or resilience (recovery from perturbation). Yet theory^{5,6}, laboratory experiments^{7–11} and indirect experimental evidence^{12–14} strongly suggest that diversity and stability are related. Here we report results from a field experiment with factorially crossed perturbation and diversity manipulations. We simulated drought perturbation on constructed grassland ecosystems containing 1, 2, 4, 8 or 32 plant species. Under unperturbed conditions, the species-poor systems achieved lower biomass production than the species-rich systems. However, the species-poor systems were more resistant to perturbation than the species-rich systems. The species-poor systems also showed a larger initial resilience following perturbation, although the original relationship between diversity and productivity was fully restored after 1 year. Our results confirm that biodiversity increases biomass production, but they also point to the fact that such diversity–production associations may lead to an inverse relationship between biodiversity and the stability of ecosystem functioning.

Ecosystems may resist change in functioning in the face of perturbation or, if change occurs, exhibit resilience by returning to their original state after perturbation¹⁵. According to the insurance hypothesis of biodiversity, resistance and resilience should increase with species richness, because a greater number of species can express a greater range of responses to environmental perturbation^{6,16}. This increases the likelihood that some species with previously low performance will increase their performance and compensate for others. However, responses of an ecosystem to perturbation may depend on the form of the relationship between the response variable and biodiversity. Autotrophic biomass production (which we will refer to as ‘biomass production’) is an important response variable that has been shown to be associated with plant species richness in experimental grassland communities^{1,3}. Given this diversity-dependence in production, there are five possible outcomes of perturbation experiments (see Fig. 1). Increasing resistance with increasing diversity (upper panel in Fig. 1b), as predicted by the insurance hypothesis, implies that the relationship between diversity and production has a larger slope under perturbed conditions than under unperturbed conditions (upper panels in Fig. 1a). This is more likely if the slope under unperturbed conditions is flat¹¹ or even negative^{10,13}. If the slope is positive, which corresponds to the situation observed in grassland experiments, the post-perturbation slope is likely to be less positive or flat because species-poor systems already have reduced biomass production under unperturbed conditions and may therefore be less affected by perturbation than species-rich systems (lower panels in Fig. 1a). In this case, resistance decreases with increasing diversity (lower panel in Fig. 1b). Analogous arguments apply to resilience. In this experiment, we test the diversity-dependence of ecosystem resistance and resilience by exposing experimental grassland communities to drought perturbation.

We used experimental communities of characteristic grassland

species where we had observed a positive relationship between biodiversity and biomass production under unperturbed conditions (the Swiss site of the BIODDEPTH experiment^{3,17,18}). After 4 years, we divided each plot, imposed an experimental drought perturbation on subplots of 1 m² and compared the above-ground biomass production in perturbed subplots with that of unperturbed control subplots. At the time of the drought treatment, the actual species richness had reached average levels of 1, 2, 4, 7 and 22 species per square metre. Replicate communities with different species combinations permitted additional contrasts to be made between plots with and without particular functional groups or species^{17,18}. The drought perturbation consisted of a transparent, height-adjustable polycarbonate roof raised above the vegetation, with a tube attached to drain the collected rainwater outside the plots. The roofs remained in place for 8 weeks, from 20 July to 18 September 1998. Climatic variables other than the amount of precipitation varied little between perturbation and control plots, but light intensity was reduced by 15% in the perturbed subplots.

The species-rich systems produced more above-ground biomass than the species-poor systems, both under unperturbed and perturbed conditions ($P < 0.001$ and $P < 0.01$, respectively, for log species richness effect tested separately for unperturbed and perturbed subplots) but the relationship was less steep in the perturbed plots (Fig. 2a, Table 1 ‘Biomass production during drought period’). Compared with pre-drought levels, diverse systems showed a greater reduction in biomass production under perturbation than did species-poor systems (Fig. 2b, Table 1 ‘Resistance’). This result contrasts with the prediction of the insurance hypothesis. Rather, it supports our predicted outcome that resistance should decrease with increasing diversity in cases where diversity–production relationships are positive under unperturbed conditions (outcome iv in Fig. 1a).

Of the alternative mechanisms that can explain these results, our

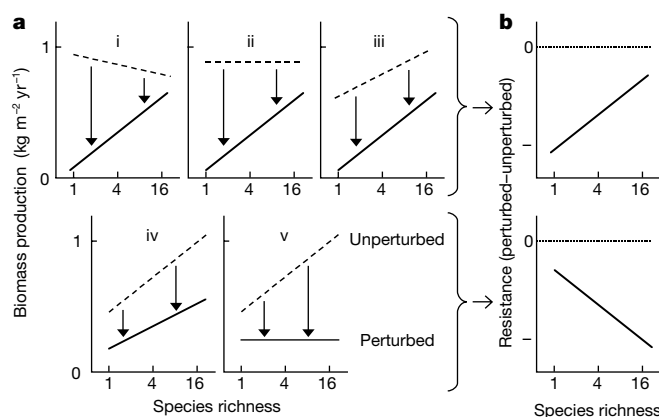


Figure 1 Possible effects of perturbation on relationships between diversity and ecosystem functioning. (Diversity here is the number of plant species on a logarithmic scale and ecosystem functioning is biomass production). **a**, Five possible outcomes in which perturbation leads to diversity-dependent reductions between unperturbed (dashed lines) and perturbed (solid lines) ecosystems. **b**, Resistance, measured as the difference between perturbed and unperturbed ecosystems (solid line; dotted reference line refers to zero difference, that is, full resistance), increases (upper panel) or decreases (lower panel) with increasing diversity, depending on the outcomes shown in **a**. Evidence that diversity increases ecosystem resistance so far comes from observations of outcomes i (refs 10, 13) and ii (ref. 11), cases in which diversity did not increase ecosystem functioning under unperturbed conditions. If diversity already has a positive effect on ecosystem functioning under unperturbed conditions, outcomes leading to positive diversity–resistance relationships (iii) may be less likely than outcomes leading to negative diversity–resistance relationships (iv, v). Our experiments demonstrate this (outcome iv), but at the same time refute the most extreme outcome (v), in which the perturbation is so severe that it reduces ecosystem functioning to a constant, diversity-independent level.

findings best support a perturbation-induced reduction in niche complementarity among plant species that probably accounts for diversity-dependent production in this system⁴. The occurrence or response of particular species in the experimental ecosystems also affected resistance, but these effects were independent of diversity. The presence of a particular functional group or species may render a community as a whole more or less resistant. When species richness was replaced by a variable coding for the presence or absence of legumes, it became apparent that plots containing legumes suffered greater reductions in biomass production under perturbed conditions than did plots without legumes ($P < 0.05$ for legume $\times$ drought interaction). A similar analysis revealed that plots containing *Poa pratensis* suffered less than plots without it ($P < 0.05$ for *P. pratensis* $\times$ drought interaction). A particular species may be more or less sensitive to perturbation independently of the community context. Thus, whereas drought reduced biomass production of the grasses *Arrhenatherum elatius*, *Dactylis glomerata* and *Trisetum flavescens* more than threefold ($P < 0.05$ for main effect of drought), the biomass production of the legumes *Lotus corniculatus* and *Trifolium pratense* and the non-leguminous forbs *Knautia arvensis* and *Plantago lanceolata* was reduced less than twofold ($P > 0.1$ for main effect of drought). These varying responses of individual species to perturbation led to decreased species evenness in perturbed subplots when compared with unperturbed subplots ($P < 0.05$ for main effect of drought). However, because species with different responses occurred at all diversity levels, the decrease in evenness due to perturbation was not affected by diversity.

We evaluated the resilience of the experimental grassland ecosystems 9 and 12 months after the drought perturbation by calculating post- to pre-drought biomass production ratios (September–June

and June–September time intervals, respectively). The recovery from perturbation was fast. At 9 months, the species-poor systems had already produced more above-ground biomass in the previously perturbed subplots than in the unperturbed subplots (Table 1 ‘Biomass production early after drought’). In the later evaluation, this was the case also for species-rich systems (Fig. 2c, Table 1 ‘Biomass production late after drought’). Thus, the initial resilience response was faster for low than for high diversity (Table 1 ‘Early resilience’), in contrast, once again, to the prediction of the insurance hypothesis and in line with our observations for the resistance response. However, species richness did not significantly influence the later resilience response (Fig. 2d, Table 1 ‘Late resilience’).

The diversity-dependent early resilience was not related to the occurrence of particular species at low diversity levels. However, the generally high resilience of our experimental systems measured after the drought perturbation was in part caused by strong increases in the biomass production of the grasses *P. pratensis*, *D. glomerata* ($P < 0.05$ for main effect of drought at both post-drought harvests) and *A. elatius* ($P < 0.05$ for main effect of drought only at the later post-drought harvest). The marked increase in *D. glomerata* and *A. elatius*, which were also among the least resistant species, was still not enough to restore the evenness of species abundances in perturbed plots to the high level measured in unperturbed plots ($P < 0.05$ for main effect of drought). This indicates that despite the full recovery made in biomass production, the responses we were recording were still transient, and even longer periods of observation would be required to show further compensatory changes.

The pre-perturbation effects of diversity have not usually been considered in theoretical work on diversity–stability relationships. Our results show that incorporating these effects can alter predictions. As we suggest in Fig. 1, the immediate effect of a drought perturbation is less pronounced in species-poor systems than in species-rich systems if biodiversity loss already reduces ecosystem functioning under unperturbed conditions. Although the insurance hypothesis did not work in the expected way, explaining increased resistance and resilience with increasing diversity, it is still con-

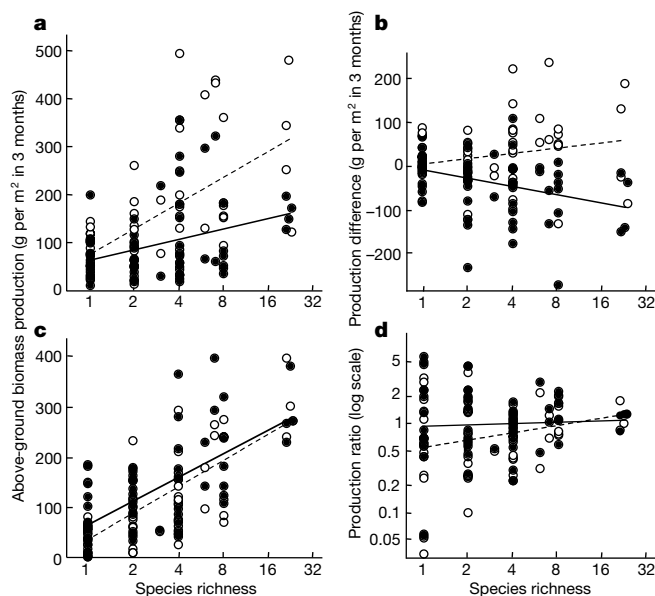


Figure 2 Effects of drought perturbation on the relationship between species richness and production. (Species richness is plotted on a logarithmic scale.) Solid lines and filled circles refer to subplots with experimental drought, and dashed lines and open circles refer to control subplots ($n = 120$). **a**, Drought perturbation reduced the slope of the diversity–production relationship (September 1998 harvest). **b**, Consequently, the difference between drought and pre-drought biomass production (September 1998–September 1997 harvests) is more negative at high than at low species richness, indicating decreasing resistance with increasing diversity. **c**, One year after the drought perturbation, the positive relationship between diversity and biomass production is fully restored (September 1999 harvest). **d**, Consequently, the ratio between pre-drought and post-drought biomass production (September 1997 / September 1999 harvests) is the same for high and low species richness, indicating that late resilience is no longer diversity-dependent (for early resilience, see the appropriate entries in Table 1).

Table 1 Perturbation effects on the diversity–production relationship

Mean in 8-species mixtures			Slope		
Control subplot	Drought subplot	Significance of drought effect	Control subplot	Drought subplot	Significance of slope difference
Biomass production during drought period*					
236.0 $\pm$ 19.9	128.2 $\pm$ 15.1	$P < 0.001$	53.6 $\pm$ 10.2	21.6 $\pm$ 7.7	$P < 0.001$
Resistance†					
40.6 $\pm$ 14.4	–67.2 $\pm$ 13.8	$P < 0.001$	12.7 $\pm$ 7.4	–19.3 $\pm$ 7.1	$P < 0.001$
Biomass production early after drought‡					
461.6 $\pm$ 31.5	462.4 $\pm$ 34.5	$P < 0.05$	96.4 $\pm$ 16.2	58.3 $\pm$ 17.7	$P < 0.05$
Early resilience§					
0.16 $\pm$ 0.14	0.20 $\pm$ 0.16	$P < 0.01$	–0.04 $\pm$ 0.07	–0.15 $\pm$ 0.08	$P < 0.05$
Biomass production late after drought					
189.1 $\pm$ 12.5	204.0 $\pm$ 15.0	$P < 0.01$	52.0 $\pm$ 6.4	46.8 $\pm$ 7.7	n.s.
Late resilience¶					
–0.04 $\pm$ 0.18	0.19 $\pm$ 0.09	$P < 0.05$	0.19 $\pm$ 0.09	0.04 $\pm$ 0.10	n.s.

Effects of drought in experimental grassland systems on the relationship between species richness (log₂ scale) and above-ground biomass production (g m^{–2}), resistance, and resilience. Estimates $\pm$ s.d. at the richness level of 8 species and slopes $\pm$ s.d. in the first, second, fifth and sixth rows refer to the regression lines in Fig. 2a–d. We calculated significance levels from multiple regression analyses (main effect of drought evaluated at mean actual species richness).

*18 June to 18 September 1998.

†Production difference drought minus pre-drought period ((18 June to 18 September 1998) – (20 June to 19 September 1997)).

‡18 September 1998 to 21 June 1999.

§Natural logarithm of production ratio early post-drought divided by pre-drought ((18 September 1998 to 21 June 1999)/(20 June to 19 September 1997)).

||21 June to 17 September 1999.

¶Natural logarithm of production ratio late post-drought divided by pre-drought ((21 June to 17 September 1999) / (20 June to 19 September 1997)).

ceivable that it worked in the opposite way: if species-rich systems have a greater chance of including species growing well under unperturbed conditions, they may also have a greater chance of losing this growth potential under perturbation if the two are negatively correlated. This offers an explanation of why more productive systems could suffer more from perturbation than those that are less productive¹⁹. Indeed, the biomass production of our perturbed systems was a more or less constant proportion of that of the unperturbed systems. When we calculated these proportions with the data presented in Fig. 2a or as ratios of drought/pre-drought biomass production (instead of the differences represented in Fig. 2b), they did not decrease significantly with increasing species richness ($P > 0.05$).

Although ecosystem resistance and resilience decreased with increasing diversity in our experiments, the relationship between diversity and biomass production was still positive, even under perturbed conditions. This shows that the perturbation was not strong enough to remove the positive diversity-dependence of ecosystem functioning altogether, and it is an argument for protecting diversity where maintaining high levels of production is desirable, even if systems with fewer species are more resistant or resilient. It is also possible that systems in which niche complementarity is not the mechanism for diversity dependence (for example, sampling or facilitation are more important⁴), in systems where multiple perturbations may exist (for example, drought, fire, insect outbreaks), and where higher trophic levels rather than diversity may govern production, different relationships between stability and diversity might be obtained. Independent of these possibilities, however, our study points to a potential trade-off between production and stability in systems where production is diversity-dependent. This trade-off may be important in understanding the ecosystem consequences of changing biodiversity. □

Methods

Biodiversity and drought experiment

The experiment was carried out at the Swiss site of the BIODDEPTH project^{3,17}, which was designed to study the influence of biodiversity loss in grassland on ecosystem processes. Two replicate blocks, each with 32 plots measuring 8 m × 2 m, were sown with different mixtures of 1, 2, 4, 8 or 32 plant species in spring 1995. The number of plant functional groups (grasses, legumes, non-leguminous forbs) ranged from one (in monocultures and in 2-, 4- and 8-species mixtures) to three (in 4-, 8- and 32-species mixtures)¹⁷. For every species and functional-group richness level, several different species compositions were used to reduce possible confusion between number and compositional effects. Similar total densities of about 500 seedlings per square metre were achieved by adjusting seed numbers in relation to germination rates and to proportions in mixtures¹⁷. The plots were mown twice during the vegetation period (June and September). To maintain the original diversity treatments, all plots were regularly weeded by hand. After 4 years, very few species had been lost from the mixtures, and four monoculture plots (*Trifolium repens* and *T. pratense*) had been destroyed by fungal infection.

After the June harvest in 1998, we applied a drought treatment within each plot and left it in position from 20 July–18 September 1998. The treatment consisted of height-adjustable, 1 m × 1 m polycarbonate roofs installed just above the vegetation (initial height 25 cm on one side, 30 cm on the other side). As the vegetation grew taller, the height of the roofs was increased twice by 10 cm in all plots. A 20-cm-deep vertical cut into the soil around the sheltered area prevented water uptake by plants with side roots. Hoses fixed to the lower end of the roofs directed the rainfall 1 m away from the plots. A 1 m × 1 m area adjacent to the roofed area was used as control. The roofs reduced light intensity by 15% and volumetric soil moisture by 13.5–27.5% with no difference between species richness levels ($P > 0.2$). Over the 8-week drought interval, the control subplots received 232 mm of natural rain (automatic weather station WS01, Delta-T Devices, Burwell, UK).

We measured the above-ground plant biomass production between the following harvest dates: 20 June and 19 September 1997, 18 June and 18 September 1998, 21 June and 17 September 1999. At each harvest, we cut the vegetation 5 cm above the ground, sorted the biomass samples from a 0.1-m² area into species and dried and weighed these separately.

Data analysis

We analysed the data resulting from the different harvests separately, because they reflected different growth intervals (3 months for the September harvests and 9 months for the June harvests) and different phases of drought response, that is, resistance (September 1998 harvest) and resilience reactions (1999 harvests). The dependent variables were (1) total above-ground biomass production and their differences (resistance) or ratios (resilience) over time, (2) evenness of partial biomass contributions of individual species to total

biomass (calculated from the Shannon diversity index H as $H/H_{\max}$; ref. 18), and (3) biomass production of individual species. We corrected the production of individual species for the decreasing proportion with which they occurred in mixtures of increasing diversity to obtain biomass production 'per seed sown'. We analysed evenness only for mixture plots (because by definition evenness is 1 for monocultures) and biomass production of individual species only for sets of plots in which they occurred. We used the logarithmic transformation for the analysis of resilience ratios in order to obtain constant variances.

We calculated the regression parameters listed in Table 1 by linear regression with log species richness as continuous explanatory variable and drought perturbation as grouping variable. For the statistical tests reported in text and Table 1 (P level), we used the analysis of variance (ANOVA) approach to multiple linear regression as implemented in Genstat software²⁰. This was necessary in order to test the number effect of species richness against the identity effect of species composition rather than against the plot variation. The full model included positional terms (block as grouping variable and distance from southern edge within blocks as continuous explanatory variable), diversity terms (log species richness as continuous explanatory variable and residual species richness and functional-group richness as grouping variables) and compositional terms (species composition as grouping variable, decomposed into contrasts for the presence/absence of particular functional groups or species) at the plot level^{17,18}, and the drought term (as grouping variable) and interactions of the diversity and compositional terms with the drought term at the sub-plot level. From this full model, we selected reduced best models using AIC values²¹. On the basis of AIC values, original species richness was replaced by actual species richness in all models the results of which are presented here. Actual species richness was measured before the drought treatment in June 1998 in subplots of 1 m².

Received 24 September; accepted 12 December 2001.

1. Tilman, D., Wedin, D. & Knops, J. M. H. Productivity and sustainability influenced by biodiversity in grassland ecosystems. *Nature* **379**, 718–720 (1996).
2. Tilman, D. *et al.* The influence of functional diversity and composition on ecosystem processes. *Science* **277**, 1300–1302 (1997).
3. Hector, A. *et al.* Plant diversity and productivity experiments in European grasslands. *Science* **286**, 1123–1127 (1999).
4. Loreau, M. *et al.* Biodiversity and ecosystem functioning: current knowledge and future challenges. *Science* **294**, 804–808 (2001).
5. May, R. M. *Stability and Complexity in Model Ecosystems* (Princeton Univ. Press, Princeton, 1973).
6. Yachi, S. & Loreau, M. Biodiversity and ecosystem productivity in a fluctuating environment: the insurance hypothesis. *Proc. Natl Acad. Sci. USA* **96**, 1463–1468 (1999).
7. McGrady-Steed, J., Harris, P. M. & Morin, P. J. Biodiversity regulates ecosystem predictability. *Nature* **390**, 162–165 (1997).
8. Naeem, S. & Li, S. Biodiversity enhances ecosystem reliability. *Nature* **390**, 507–509 (1997).
9. Petchey, O. L., McPhearson, P. T., Casey, T. M. & Morin, P. J. Environmental warming alters food-web structure and ecosystem function. *Nature* **402**, 69–72 (1999).
10. Griffiths, B. S. *et al.* Ecosystem response of pasture soil communities to fumigation-induced microbial diversity reductions: an examination of the biodiversity–ecosystem function relationship. *Oikos* **90**, 279–294 (2000).
11. Mulder, C. P. H., Uliassi, D. D. & Doak, D. F. Physical stress and diversity–productivity relationships: The role of positive interactions. *Proc. Natl Acad. Sci. USA* **98**, 6704–6708 (2001).
12. Dodd, M. E., Silvertown, J., McConway, K., Potts, J. & Crawley, M. Stability in the plant-communities of the Park Grass experiment—the relationships between species richness, soil-pH and biomass variability. *Phil. Trans. R. Soc. Lond. B* **346**, 185–193 (1994).
13. Tilman, D. Biodiversity: population versus ecosystem stability. *Ecology* **77**, 350–363 (1996).
14. Sankaran, M. & McNaughton, S. J. Determinants of biodiversity regulate compositional stability of communities. *Nature* **401**, 691–693 (1999).
15. Herbert, D. A., Fownes, J. H. & Vitousek, P. M. Hurricane damage to a Hawaiian forest: nutrient supply rate affects resistance and resilience. *Ecology* **80**, 908–920 (1999).
16. Lawton, J. H. & Brown, V. K. in *Biodiversity and Ecosystem Function* (eds Schulze, E.-D. & Mooney, H. A.) 255–270 (Springer, Berlin, 1993).
17. Diemer, M., Joshi, J., Körner, C., Schmid, B. & Spehn, E. An experimental protocol to assess the effects of plant diversity on ecosystem functioning utilized in a European research network. *Bull. Geobot. Inst. ETH* **63**, 95–107 (1997).
18. Stephan, A., Meyer, A. & Schmid, B. Plant diversity affects culturable soil bacteria in experimental grassland communities. *J. Ecol.* **88**, 988–998 (2000).
19. Huston, M. A. Hidden treatments in ecological experiments: re-evaluating the ecosystem function of biodiversity. *Oecologia* **110**, 449–460 (1997).
20. Payne, R. W. *et al.* *GENSTAT 5 Reference Manual* (Clarendon, Oxford, 1993).
21. Burnham, K. P. & Anderson, D. R. *Model Selection and Inference: a Practical Information-Theoretic Approach* (Springer, New York, 1998).

Acknowledgements

We thank M. Diemer for help with the experimental set-up, M. Fischer, A. Hector, M. Loreau, P. Oertli-Barnett and O. L. Petchey for suggestions on earlier drafts of this manuscript, and numerous helpers in the field for assistance. This project was supported by the Swiss Federal Office for Education and Science as part of the EU-funded BIODDEPTH project.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to B.S. (e-mail: bschmid@uwinst.unizh.ch).

Chimaeric sounds reveal dichotomies in auditory perception

Zachary M. Smith^{*†‡}, Bertrand Delgutte^{*†‡} & Andrew J. Oxenham^{†‡}

^{*} Eaton-Peabody Laboratory, Massachusetts Eye & Ear Infirmary, Boston, Massachusetts 02114, USA

[†] Research Laboratory of Electronics; and [‡] Speech and Hearing Bioscience and Technology Program, Harvard-MIT Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

By Fourier's theorem¹, signals can be decomposed into a sum of sinusoids of different frequencies. This is especially relevant for hearing, because the inner ear performs a form of mechanical Fourier transform by mapping frequencies along the length of the cochlear partition. An alternative signal decomposition, originated by Hilbert², is to factor a signal into the product of a slowly varying envelope and a rapidly varying fine time structure. Neurons in the auditory brainstem^{3–6} sensitive to these features have been found in mammalian physiological studies. To investigate the relative perceptual importance of envelope and fine structure, we synthesized stimuli that we call 'auditory chimaeras', which have the envelope of one sound and the fine structure of another. Here we show that the envelope is most important for speech reception, and the fine structure is most important for pitch perception and sound localization. When the two features are in conflict, the sound of speech is heard at a location determined by the fine structure, but the words are identified according to the envelope. This finding reveals a possible acoustic basis for the hypothesized 'what' and 'where' pathways in the auditory cortex^{7–10}.

Combinations of features from different sounds have been used in the past to produce new, hybrid sounds for use in electronic music^{11,12}. Our aim in combining features from different sounds was to study the perceptual relevance of the envelope and fine structure in different acoustic situations. To synthesize auditory chimaeras, two sound waveforms are used as inputs. A bank of band-pass filters is used to split each sound into 1 to 64 complementary frequency bands spanning the range 80–8,820 Hz. Such splitting into frequency bands resembles the Fourier analysis performed by the cochlea and by processors for cochlear implants. The output of each filter is factored into its envelope and fine structure using the Hilbert transform (see Methods). The envelope of each filter output from the first sound is then multiplied by the fine structure of the corresponding filter output from the second sound. These products are finally summed over all frequency bands to produce an auditory chimaera that is made up of the envelope of the first sound and the fine structure of the second sound in each band. The primary variable in this study is the number of frequency bands, which is inversely related to the width of each band. A block diagram of chimaera synthesis is shown in Fig. 1 with example waveforms for a single frequency band. For an audio demonstration of auditory chimaeras, see ref. 13.

Speech is a robust signal that can be perturbed in many different ways while remaining intelligible^{14,15}. Speech chimaeras were created by combining either a speech sentence and noise or by combining two separate speech sentences. The speech material comprised sentences from the Hearing in Noise Test (HINT)¹⁶. Speech–noise chimaeras were synthesized from individual HINT sentences and spectrally matched noise. These chimaeras contain speech information in either their envelope or their fine structure. Speech–speech chimaeras were synthesized from two different HINT sentences of similar duration. The envelope of each speech–speech chimaera contains information about one utterance; its fine structure contains information about another.

Listening tests with speech–noise chimaeras showed that speech reception is highly dependent on the number of frequency bands used for synthesis (Fig. 2). When speech information is contained solely in the envelope, speech reception is poor with one or two frequency bands and improves as the number of bands increases. Good performance (>85% word recognition) is achieved with as few as four frequency bands, consistent with previous findings that bands of noise modulated by speech envelope can produce good speech reception with very limited spectral information¹⁷. In contrast, when speech information is only contained in the fine structure, speech reception is generally better with fewer frequency bands. The best performance is achieved with two bands; performance then deteriorates as the number of bands increases until, with eight or more bands, there is essentially no speech reception. Good performance with one and two frequency bands of fine structure is consistent with previous findings that peak-clipping (which flattens out the envelope) does not severely degrade speech reception¹⁴. Poorer performance with increasing numbers of bands is consistent with the auditory system's insensitivity to the fine structure of critical-band signals at high frequencies¹⁸.

Speech–speech chimaeras measure the relative salience of the speech information transmitted through the envelope and fine structure when the two types of information are conflicting. Even though speech–speech chimaeras are constructed with two distinct

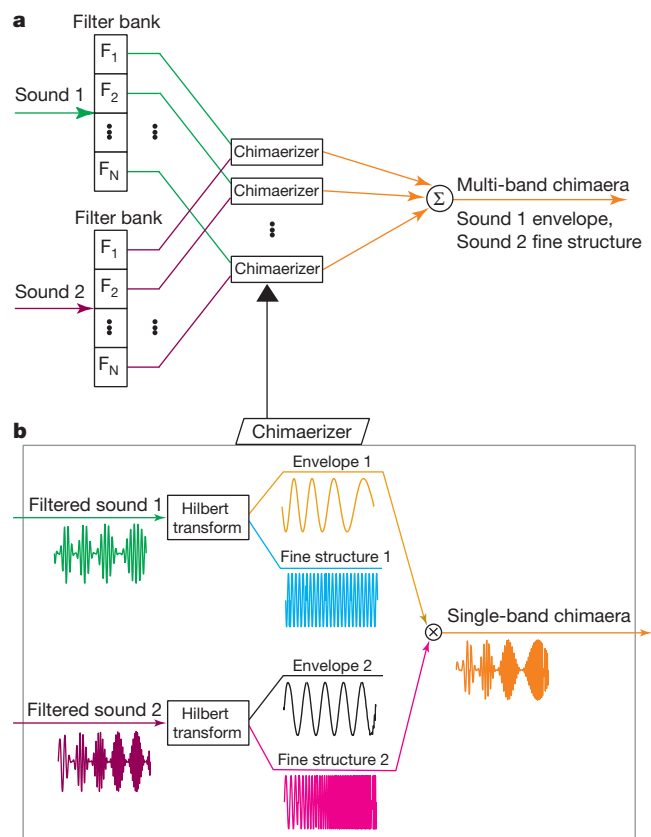


Figure 1 Auditory chimaera synthesis. **a**, Two sounds are used as input. Each sound is split into N complementary frequency bands with a perfect-reconstruction filter bank. Filtered signals from matching frequency bands are processed through a chimaerizer, which exchanges the envelope and the fine time structure of the two signals, producing a single-band chimaera. Partial chimaeras are summed over all frequency bands to produce a multi-band chimaera. **b**, Example waveforms within a chimaerizer, where band-limited input signals are factored into their envelope and fine structure using the Hilbert transform. A single-band auditory chimaera is made from the product of envelope 1 and fine structure 2.

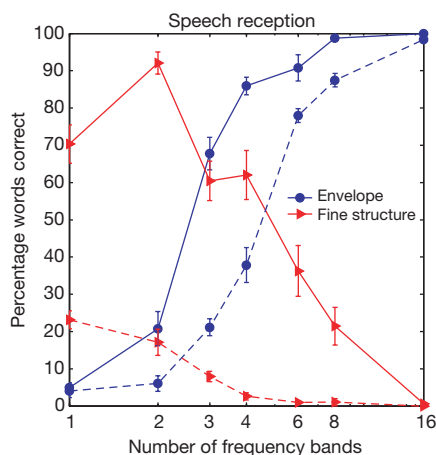


Figure 2 Speech reception of sentences in the envelope and fine structure of auditory chimaeras. Speech–noise chimaeras (solid lines) only contain speech information in either the envelope or the fine structure. Speech–speech chimaeras (dashed lines) have conflicting speech information in the envelope and the fine structure.

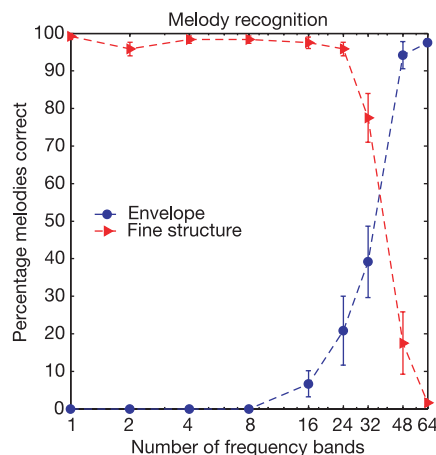


Figure 3 Recognition of melodies in the envelope and in the fine structure of auditory chimaeras. Melody–melody chimaeras contain conflicting melodies in the envelope and fine structure.

utterances, listeners almost invariably heard words from only one of the two sentences. In general, the speech information contained in the envelope dominated the information contained in the fine structure. Speech reception based on the fine structure was much poorer with speech–speech chimaeras than with speech–noise chimaeras, and was only above chance with one and two frequency bands. Speech reception based on envelope information was also degraded, but not as severely, and performance still exceeded 80% with eight or more frequency bands. Thus, envelope information in speech is more resistant to conflicting fine-structure information from another sentence than vice versa.

We used a melody recognition task to assess pitch perception of complex harmonic sounds. For this purpose, we synthesized chimaeras based on two different melodies, one in the envelope and the other in the fine structure. Melody–melody chimaeras show a reversal in the relative importance of envelope and fine structure when compared to speech–speech chimaeras (Fig. 3). Listeners always heard the melody based on the fine structure with up to 32 frequency bands. With 48 and 64 frequency bands, however, they identified the envelope-based melody more often than they did the melody based on fine structure. In responses to these melody–melody chimaeras, subjects sometimes reported hearing two melodies and often picked the two melodies represented in the envelope and the fine structure, respectively. This can be seen in the data with 16 to 48 frequency bands where scores based on envelope and fine structure add up to more than 100% correct.

The crossover point, where the envelope begins to dominate over the fine structure, occurs for a much higher number of frequency bands (about 40) for melody–melody chimaeras than it does for speech–speech and speech–noise chimaeras, further suggesting that speech reception depends primarily on envelope information in broad frequency bands. For melody recognition, this crossover occurs approximately when the bandwidths of the bandpass filters become narrower than the critical bandwidths. For such narrow bandwidths, precise information about the frequency of each spectral component, and hence the overall pitch, is available in the spectral distribution of the envelopes.

Sound localization in the horizontal plane is based on interaural differences in time and level. Interaural time differences (ITD) are the dominant cue for low-frequency sounds such as speech¹⁹. A delay of 700 μ s was introduced into either the right or left channel of each HINT sentence to create ITDs that would produce completely lateralized sound images. To synthesize dichotic chimaeras, a sentence with an ITD pointing to the right was combined with a

sentence having an ITD pointing to the left to produce a chimaera with its envelope information pointing to one side and its fine structure pointing to the other side. Two types of dichotic chimaeras were constructed, one using the same sentence for both the envelope and fine structure, and the other using different sentences. Lateralization of dichotic chimaeras was always based on the ITD of the fine structure (Fig. 4), consistent with results using non-speech stimuli^{20,21}. Chimaeras synthesized with a small number of frequency bands were difficult to lateralize, but lateralization improved with increasing number of bands. Dichotic chimaeras based on the same sentence in the envelope and fine structure were more easily lateralized than those based on two different sentences.

When dichotic chimaeras based on different sentences were presented, listeners were asked to pick which of the two sentences they heard in addition to reporting the lateral position of the sound image. Consistent with our results for speech–speech chimaeras, subjects most often heard the sentence based on the fine structure with one and two frequency bands, whereas they heard the sentence based on the envelope for four or more bands. With eight or more frequency bands, subjects clearly identified the sentence based on the envelope but lateralized the speech to the side to which the fine structure was pointing. Thus the fine structure determines ‘where’ the sound is heard, whereas the envelope determines ‘what’ sentence is heard. In this respect, auditory chimaeras are consistent with evidence for separate ‘where’ and ‘what’ pathways in the auditory cortex^{7–10}.

The Hilbert transform provides a mathematically rigorous definition of envelope and fine structure free of arbitrary parameters. The squared Hilbert envelope contains frequencies up to the bandwidth of the original signal. Thus, for low numbers of frequency bands (less than about six), and hence large filter bandwidths, the fluctuations in the envelope can become rather rapid, making the functional distinction between fine structure and envelope based on fluctuation rate difficult. One of our most important findings is that the perceptual importance of the envelope increases with the number of frequency bands, while that of the fine structure diminishes (Figs 2 and 3). Had we used a different technique (such as rectification followed by low-pass filtering) to extract a smoother envelope, the smooth envelope would have contained even less information for small numbers of frequency bands, and therefore its perceptual importance would probably have been even smaller. Furthermore, a previous study¹⁷ has indicated that eliminating all rapid envelope fluctuations from bands of noise modulated by speech envelope has little or no effect on speech

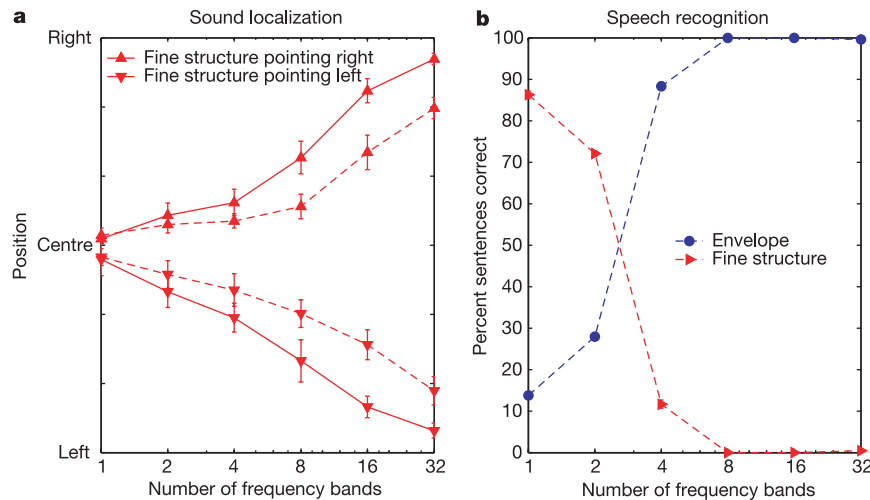


Figure 4 'What' and 'where' for dichotic chimaeras are determined by different cues. **a**, Dichotic chimaeras constructed with the same sentence are shown with solid lines and those made from different sentence are shown with dashed lines. **b**, Subjects heard the

reception. Thus, for our purposes, it is valid to consider the envelope as slowly varying relative to the fine structure.

Cochlear implants are prosthetic devices that seek to restore hearing in the profoundly deaf by stimulating the auditory nerve via electrodes inserted into the cochlea. Current processors for cochlear implants discard the fine time structure, and present only about six to eight bands of envelope information²². Our results suggest that modifying cochlear implant processors to deliver fine-structure information may improve patients' pitch perception and sensitivity to ITD. Better pitch perception should benefit music appreciation. It should also help convey prosody cues in speech and may enhance speech reception among speakers of tonal languages, such as Mandarin Chinese, where pitch is used to distinguish different words. Better ITD sensitivity may help the increasing number of patients with bilateral cochlear implants in taking advantage of binaural cues that normal-hearing listeners use to distinguish speech among competing sound sources. Strategies for improving the representation of fine structure in cochlear implants have been proposed²³ and supported by single-unit data²⁴. □

Methods

Stimulus synthesis

The perfect-reconstruction digital filter banks used for chimaera synthesis spanned the range 80–8,820 Hz, spaced in equal steps along the cochlear frequency map²⁵ (nearly logarithmic frequency spacing). For example, with six bands, the cutoff frequencies were 80, 260, 600, 1,240, 2,420, 4,650 and 8,820 Hz. The transition over which adjacent filters overlap significantly was 25% of the bandwidth of the narrowest filter in the bank (the lowest in frequency). Thus, for the six band case, each filter transition was 45 Hz wide.

To compute the envelope and fine structure in each band, we used the analytic signal²⁶ $s(t) = s_r(t) + i s_i(t)$, where $s_r(t)$ is the filter output in one band, $s_i(t)$ the Hilbert transform of $s_r(t)$, and $i = \sqrt{-1}$. The Hilbert envelope is the magnitude of the analytic signal, $a(t) = \sqrt{s_r^2(t) + s_i^2(t)}$. The fine structure is $\cos \phi(t)$, where $\phi(t) = \arctan(s_i(t)/s_r(t))$ is the phase of the analytic signal. The original signal can be reconstructed as $s_r(t) = a(t) \cos \phi(t)$. In practice, the Hilbert transform was combined with the band-pass filtering operation using complex filters whose real and imaginary (subscripts r, i) parts are in quadrature²⁷.

Subjects and procedure

Six native speakers of American English with normal hearing thresholds participated in each part of the study. Five of these subjects participated in the entire series of tests. Speech reception, melody recognition, and lateralization tests were conducted separately. Within each individual experiment, the order of all conditions was randomized. Stimuli were presented in a soundproof booth through headphones at a root-mean-square sound pressure level of 67 dB.

In the speech reception experiment, subjects listened to the processed sentences and were instructed to type the words they heard into a computer. Each subject listened to a total of 273 speech chimaeras with an additional seven for training. Speech reception was measured as percentage words correct. 'The', 'a' and 'an' were not scored. When speech–

sentence in the envelope more than the sentence in the fine structure when four or more bands were used.

speech chimaeras were used, each word in a subject's response could count for either a sentence in the envelope or in the fine structure, but this condition rarely occurred in practice.

Before the melody recognition experiment, each subject selected ten melodies that he/she was familiar with. Each melody was taken from a set of 34 simple melodies with all rhythmic information removed²⁸, consisting of 16 equal-duration notes and synthesized with MIDI software that used samples of a grand piano. During the experiment, subjects selected from their own list of ten melodies which one(s) they heard on each trial. Melodies were scored as percentage correct even when subjects reported multiple melodies in a single trial without penalty for incorrect responses.

In the lateralization experiment, subjects used a seven-point scale to rate the lateral position of the sound image inside the head. This scale ranges from -3 to +3, with -3 corresponding to the left ear and +3 to the right ear. Lateralization scores were averaged for each condition. In addition, subjects had to select which of two possible sentences they heard, one choice corresponding to the envelope and the other one to the fine structure.

Received 15 October; accepted 12 December 2001.

- Fourier, J. B. J. La théorie analytique de la chaleur. *Mém. Acad. R. Sci.* **8**, 581–622 (1829).
- Hilbert, D. *Grundzüge einer allgemeinen Theorie der linearen Integralgleichungen* (Teubner, Leipzig, 1912).
- Rhode, W. S., Oertel, D. & Smith, P. H. Physiological response properties of cells labeled intracellularly with horseradish peroxidase in cat ventral cochlear nucleus. *J. Comp. Neurol.* **213**, 448–463 (1983).
- Joris, P. X. & Yin, T. C. Envelope coding in the lateral superior olive. I. Sensitivity to interaural time differences. *J. Neurophysiol.* **73**, 1043–1062 (1995).
- Yin, T. C. & Chan, J. C. Interaural time sensitivity in medial superior olive of cat. *J. Neurophysiol.* **65**, 465–488 (1990).
- Langner, G. & Schreiner, C. E. Periodicity coding in the inferior colliculus of the cat. I. Neuronal mechanisms. *J. Neurophysiol.* **60**, 1799–1822 (1988).
- Rauschecker, J. P. & Tian, B. Mechanisms and streams for processing of "what" and "where" in auditory cortex. *Proc. Natl Acad. Sci. USA* **97**, 11800–11806 (2000).
- Recanzone, G. H. Spatial processing in the auditory cortex of the macaque monkey. *Proc. Natl Acad. Sci. USA* **97**, 11829–11835 (2000).
- Tian, B., Reser, D., Durham, A., Kustov, A. & Rauschecker, J. P. Functional specialization in rhesus monkey auditory cortex. *Science* **292**, 290–293 (2001).
- Maeder, P. P. et al. Distinct pathways involved in sound recognition and localization: a human fMRI study. *NeuroImage* **14**, 802–816 (2001).
- Dodge, C. & Jerse, T. A. *Computer Music: Synthesis, Composition, and Performance* (Schirmer Books, New York, 1997).
- Depalle, P., García, G. & Rodet, X. A virtual castrato (?). *Proc. Int. Computer Music Conf., Aarhus, Denmark* 357–360 (1994).
- Shen, C., Smith, Z. M., Oxenham, A. J. & Delgutte, B. Auditory Chimera Demo; available at (<http://epl.meei.harvard.edu/~bard/chimera>) (2001).
- Licklider, J. Effect of amplitude distortion upon the intelligibility of speech. *J. Acoust. Soc. Am.* **18**, 429–434 (1946).
- Saberi, K. & Perrot, D. R. Cognitive restoration of reversed speech. *Nature* **398**, 760 (1999).
- Nilsson, M., Soli, S. D. & Sullivan, J. A. Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. *J. Acoust. Soc. Am.* **95**, 1085–1099 (1994).
- Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J. & Ekelid, M. Speech recognition with primarily temporal cues. *Science* **270**, 303–304 (1995).
- Ghizla, O. On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception. *J. Acoust. Soc. Am.* **110**, 1628–1640 (2001).
- Wightman, F. L. & Kistler, D. J. The dominant role of low-frequency interaural time differences in sound localization. *J. Acoust. Soc. Am.* **91**, 1648–1661 (1992).
- Henning, G. B. & Ashton, J. The effect of carrier and modulation frequency on lateralization based on interaural phase and interaural group delay. *Hear. Res.* **4**, 184–194 (1981).

21. Bernstein, L. R. & Trahiotis, C. Lateralization of low-frequency complex waveforms: The use of envelope-based temporal disparities. *J. Acoust. Soc. Am.* **77**, 1868–1880 (1985).
22. Wilson, B. S. *et al.* Better speech recognition with cochlear implants. *Nature* **352**, 236–238 (1991).
23. Rubinstein, J. T., Wilson, B. S., Finley, C. C. & Abbas, P. J. Pseudospontaneous activity: stochastic independence of auditory nerve fibers with electrical stimulation. *Hear. Res.* **127**, 108–118 (1999).
24. Litvak, L., Delgutte, B. & Eddington, D. Auditory nerve fiber responses to electrical stimulation: modulated and unmodulated pulse trains. *J. Acoust. Soc. Am.* **110**, 368–379 (2001).
25. Greenwood, D. D. A cochlear frequency-position function for several species—29 years later. *J. Acoust. Soc. Am.* **87**, 2592–2604 (1990).
26. Ville, J. Théorie et applications de la notion de signal analytique. *Cables Transmission* **2**, 61–74 (1948).
27. Troullos, G., Ehlig, P., Chirayil, R., Bradley, J. & Garcia, D. in *Digital Signal Processing Applications with the TMS320 Family* (ed. Papamichalis, P.) 221–330 (Texas Instruments, Dallas, 1990).
28. Hartmann, W. M. & Johnson, D. Stream segregation and peripheral channeling. *Music Percept.* **9**, 155–184 (1991).

Acknowledgements

We thank C. Shen for assistance with data analysis and J. R. Melcher and L. D. Braidia for comments on an earlier version of the manuscript. A.J.O. is currently a fellow at the Hanse Institute for Advanced Study in Delmenhorst, Germany. This work was supported by grants from the National Institutes of Health (NIDCD).

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to B.D. (e-mail: bard@epi.meei.harvard.edu).

Long-term plasticity in hippocampal place-cell representation of environmental geometry

Colin Lever*, Tom Wills*, Francesca Cacucci*, Neil Burgess*† & John O'Keefe*†

* Department of Anatomy and Developmental Biology; and † Institute of Cognitive Neuroscience, University College London, WC1E 6BT, UK

The hippocampus is widely believed to be involved in the storage or consolidation of long-term memories^{1–4}. Several reports have shown short-term changes in single hippocampal unit activity during memory and plasticity experiments^{5–12}, but there has been no experimental demonstration of long-term persistent changes in neuronal activity in any region except primary cortical areas^{13–16}. Here we report that, in rats repeatedly exposed to two differently shaped environments, the hippocampal-place-cell representations of those environments gradually and incrementally diverge; this divergence is specific to environmental shape, occurs independently of explicit reward, persists for periods of at least one month, and transfers to new enclosures of the same shape. These results indicate that place cells may be a neural substrate for long-term incidental learning, and demonstrate the long-term stability of an experience-dependent firing pattern in the hippocampal formation.

In rats, hippocampal lesions cause deficits in spatial behaviour^{2,17–20}. One of the major behavioural correlates of the firing of hippocampal pyramidal cells is the animal's location. Previous experimental and theoretical work suggests that the major determinant of the location and shape of place-cell firing fields is the distance from two or more walls in particular directions^{21,22}. This theory predicts that these cells will have related patterns of firing in enclosures of different shape. We tested this prediction (see Methods 'preliminary experiment' and Supplementary Information) and found, in each of seven rats, that place fields were

very similar on initial exposure to square and circular boxes (Fig. 1a). 73% of the cells (48/66) had 'homotopic' fields in both shapes (that is, in the same location, see Methods for definition of 'homotopic'). Other environmental manipulations, such as translation (Fig. 1b), removal (Fig. 1c), and reconfiguration of the box into shapes other than squares and circles (not shown), showed that firing patterns relate to the box walls and not to other cues in the testing arena. This finding of similarity across shapes conflicts with earlier experiments^{7,23}, performed on animals that had had considerable experience of the testing enclosures. We asked whether experience was the critical factor in producing neuronal discrimination between different shapes.

We recorded hippocampal CA1 cells from a new group of three animals during repeated exposures to different shapes of enclosure (see Methods 'main experiment'). Recording was performed during the animals' entire experience (up to three weeks) of these enclosures: unlike previous studies (for example, refs 5,6,7,8,9,10,11,12, 21, 23,24,25,26,27,28,29,30), there was no unrecorded training phase. Four boxes were used: two identical circular-walled, and two identical square-walled boxes (hereafter simply 'circle' and 'square'), all made of the same materials so that discrimination on the basis of geometry could be separated from discrimination on the basis of other differences between boxes.

We recorded on successive days monitoring the activity of the same group of neurons where possible, in some cases following individual cells for over a week. In other cases we obtained different samples from the same overall population. Either way, on the first day the firing patterns in the two shapes were similar, replicating our previous work; on later trials, the patterns diverged while those in

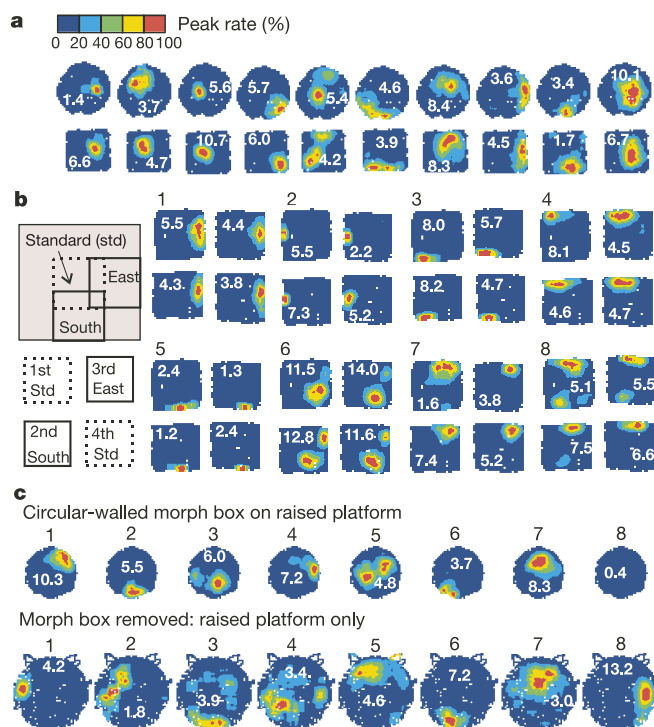


Figure 1 Similarity of spatial firing during early exposure to circular- and square-walled enclosures. **a**, Similar place fields of 10 representative simultaneously recorded CA1 neurons in circle and square. Probe trials on two different groups of cells (**b**, **c**) show that this similarity is determined by box walls rather than similar sets of background cues in the testing arena in both circle and square conditions. Box-wall translation by 40 cm (**b**) eastwards (upper right) or southwards (lower left) does not affect firing fields relative to the box frame, while box-wall removal (**c**) induces remapping. Fields with less than a 1.0-Hz peak rate are not shown.

the same shape remained similar. Figure 2 summarizes this process and also shows that the divergence endures after a delay. Across all three animals, 14/17 cells (82%) had homotopic fields on day 1 (Fig. 2a), after which remapping proceeded, but not at the same rate in all animals. Rat 1 (data shown by an asterisk) had three identical fields in the two shapes on day 1 but had completely remapped by day 5 (showing no overlap between six fields). After similar data was obtained on day 6, time-series recording from this animal was terminated. Rats 2 and 3 took considerably longer to remap with 12/14 place cells displaying 'similar' fields on day 1, 7/13 on day 5, 7/32 on day 21 (see Methods for numerical quantification of 'similar'). Examination of the animals' behaviour during these days did not reveal any systematic differences that would explain these changes

(see Supplementary Information).

The firing patterns in the square and circle on day 21 (Fig. 2d) differ either because the cell fires in one shape and not the other (top row), or because it fires in two different places in the two shapes (bottom row). To quantify these changes, we transformed the circle topologically into a square of the same area as the square box and used these transformed data to compute three different measures of disparity (see Methods). The first, rate divergence (RD), is sensitive to the difference in firing rates between boxes of different shape compared to boxes of the same shape; the second, peak divergence (PD), is sensitive to the difference between the location of firing field peaks in the two shapes compared to boxes of the same shape, and the third, field divergence (FD), is the proportion of the cells with

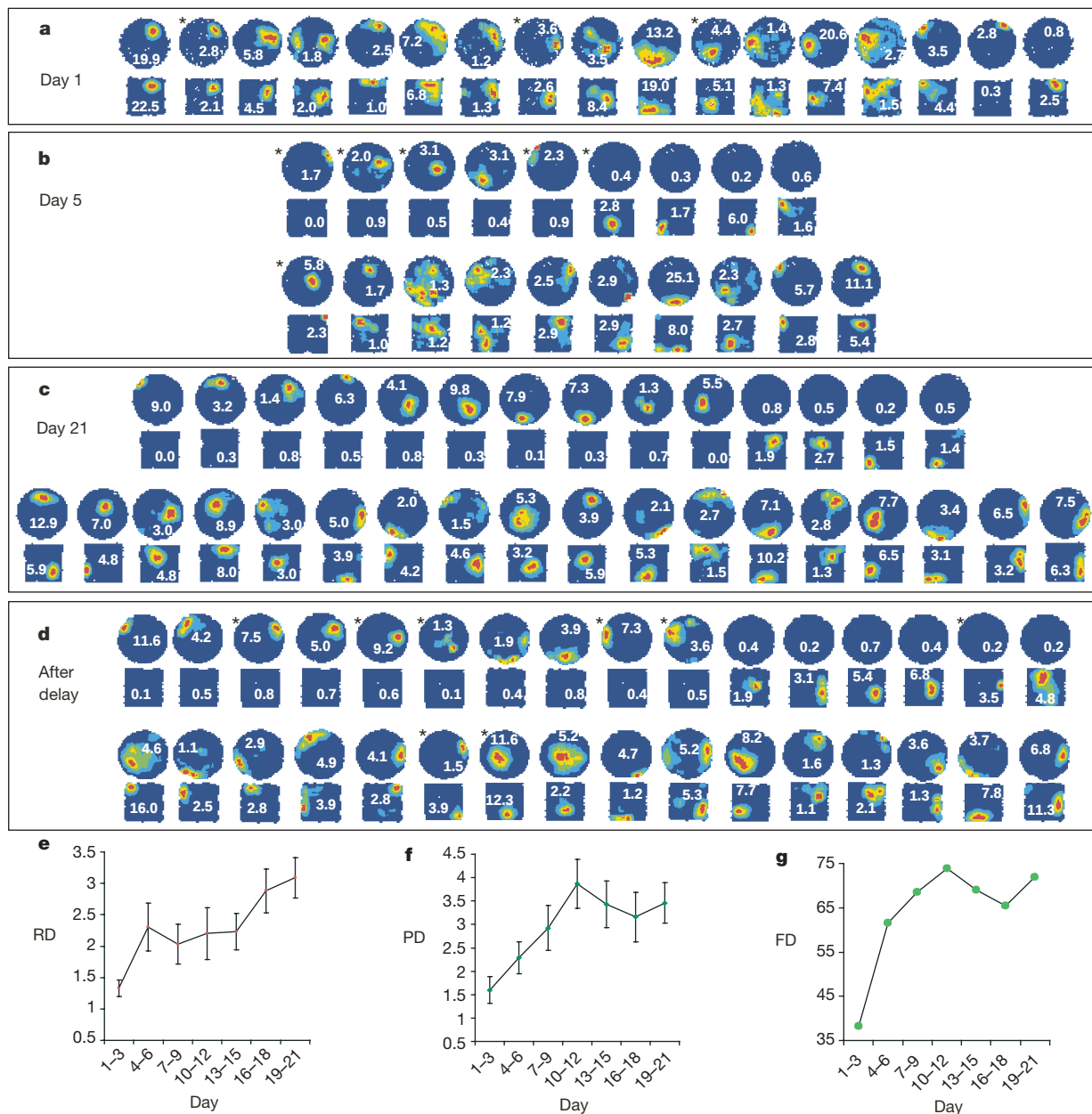


Figure 2 Repeated exposure to circle and square enclosures increases place field divergence. Similar patterns on day 1 (**a**) are followed by more divergent patterns on day 5 (**b**), and still more by day 21 (**c**). Pattern divergence is maintained after long delays (**d**). Various measures of firing pattern divergence from the slower-learning rats 2 and 3 show experience induces increases in: **e**, rate divergence (RD: estimated range 1.0 to 5.3);

f, peak divergence (PD: estimated range 1.0 to 5.3); and **g**, proportion of cells with divergent fields (FD: estimated range 29% to 90%). Graphs show mean values (error bars in **e**, **f** are $\pm$ s.e.m.) over 3-day blocks. See Supplementary Information for details. Asterisks indicate data from faster-learning Rat 1.

fields in both shapes that are not homotopic. Figure 2e–g depicts these measures in rats 2 and 3, averaged in 3-day blocks. All three measures show an increase over time (correlation between time and mean RD: $r = 0.89$, $P < 0.004$; mean PD: $r = 0.78$, $P < 0.025$; mean FD: $r = 0.70$, $P < 0.05$). The cells of rat 1 remapped primarily

by rate divergence (mean RD correlation with time: $r = 0.91$, $P < 0.01$). Reflecting these trends, the firing patterns at the end of the time series ('dayend': day 6 for rat 1, day 21 for rats 2 and 3) were significantly more divergent than on day 1: mean RD increased from 1.32 ± 0.22 to 3.59 ± 0.69 ($P < 0.002$); mean PD from 1.15 ± 0.19 to 4.00 ± 0.59 ($P < 0.0001$); FD from 18% (3/17) to 76% (16/21) (χ^2 , $P < 0.001$).

The gradual shift in the population statistics (averaged across cells and animals) could be due to abrupt transitions at different times in different cells, or to gradual changes within each cell. We were able to track the activity of several individual cells throughout their transition period and found that most diverged in an incremental fashion. Figure 3a shows three different ways in which neurons with initially homotopic fields in both boxes were observed to differentiate between them. Figure 3b–d shows one example of each. The cell in Fig. 3b develops a new subfield whose rate progressively increases over the trial while that in the old field decreases; this process repeats on the subsequent trial and the new field remains thereafter. The cell in Fig. 3c illustrates the progressive translation of field position in one box over time. The cell in Fig. 3d is an example of a common change, a progressive reduction of firing rate to below the 1-Hz threshold in one shape together with the maintenance of firing rate in the other. In each case, the divergence process can be seen to take place over several trials or days.

We asked whether this remapping was permanent. Each animal was returned to its home cage and not replaced in either environment for a period of several weeks (28, 17 and 39 days in rats 1–3 respectively) after its last exposure there. After this delay, the animals were retested in the circle and square. In all cases the majority of cells differentiated between the two boxes (see Fig. 2d), suggesting that the learning process was a permanent one. None of the 8 cells recorded from rat 1 (asterisks) have similar fields, and only 5 of the 24 cells from the slower-learning rats 2 and 3 have similar fields. The overall pattern remains highly divergent with no significant differences in our three measures between the last training day and the day of the delay test (all $P > 0.05$). In a further (additional 29 days) delay test on rat 2 after that shown in Fig. 2d, only 3/10 cells had similar firing patterns (see Supplementary Information).

It is unlikely that this shape-specific divergence in place fields could be due to the passage of time alone rather than to the rats' experience of the enclosures. Some corroboration of this was obtained in a delay test on one of the animals from the preliminary experiment. On its first day of exposure to the two boxes, 9/10 cells had similar fields (mean across-shape distance 7.2 ± 1.0 cm). This rat then had six further days exposure to the two shapes, followed by a series of probe trials in the square box only, after which there were minimal signs of remapping, 5/7 cells being similar (mean across-shape distance 7.2 ± 1.7 cm). Following a delay of 32 days without further experience of the boxes, 5/5 cells (see Supplementary Information) were similar (mean across-shape distance 5.5 ± 1.7 cm). We conclude that the post-delay pattern divergence seen in rats 1, 2 and 3 indicates long-term storage and retrieval of an experience-dependent learned pattern.

What is the perceptual basis for the acquired environmental discrimination, and is it transferable? Two wooden squares and two wooden circles were used during the learning phase, and all measures compared inter-shape with intra-shape changes, so the cells must become tuned to geometrical features. We wondered whether geometrical tuning was strong enough to override changes in non-geometrical features in a transfer test. Rats 1 to 3, trained in the wooden boxes, were now tested in the 'morph' circle and 'morph' square, two environments created from the same deformable structure (see Methods). In addition, two new rats (4, 5) were trained in the morph circle and morph square and then tested in the wooden circle and wooden square boxes. We recorded 65 cells in total (41 in the wooden-to-morph box transfer study, 24 in the

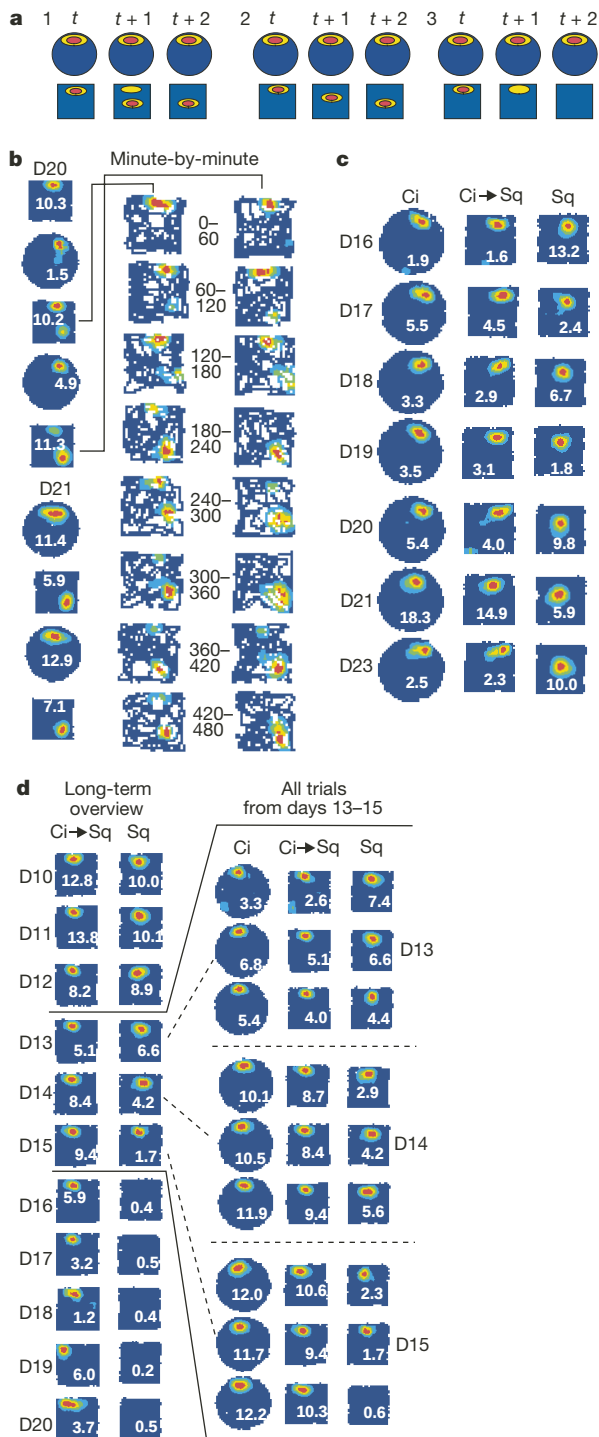


Figure 3 Schematic (a) and observed (b, c, d) incremental development of divergent firing in individual place cells. 'Ci → Sq' denotes transformed-circle firing-rate maps (see Methods). Dn denotes recording day. b, Cell develops a new subfield in square. c, Cell shifts its field in square relative to circle. d, Cell reduces its firing rate in square, relative to circle. Peak rates in square are about equal to those in circle on day 13, about 50% on day 14, less than 25% on day 15, and below the 1-Hz threshold thereafter.

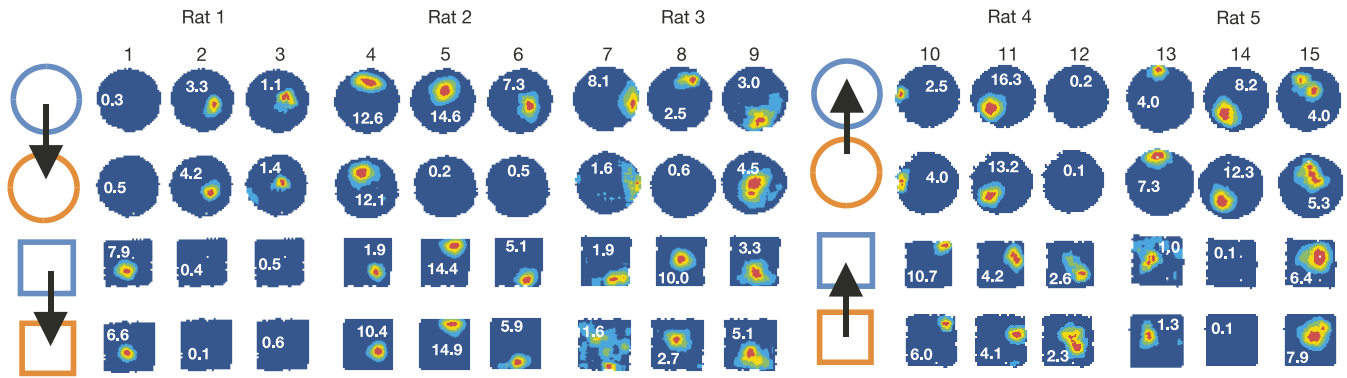


Figure 4 Shape-specific firing patterns learned in one type of enclosure were reproduced in different enclosures of similar shape. Diagrams indicate type and shape of enclosure (blue, wooden box; orange, 'morph' box). Firing-rate maps from three representative cells in each of five animals are shown in rows corresponding (top to bottom) to wooden circle,

morph circle, wooden square, morph square. Firing-rate maps of rats originally trained in the wooden boxes (rats 1, 2, 3) are shown on the left, and those of rats originally trained in the morph box (rats 4, 5) on the right.

morph-to-wooden box transfer study). Figure 4 shows three representative cells for each rat. Overall we found firing patterns to be strongly influenced by geometrical shape, despite the change in box construction and material. The mean distance between the place field peaks of cells firing in boxes of the same shape but different material was only 7.0 ± 0.6 cm ($n = 51$, $P \ll 1 \times 10^{-10}$, see the Monte Carlo simulation in Methods). The differences in cells' firing rates between trials in boxes of different material were not significantly different to the differences in firing rates between trials in the boxes of the same material ($P > 0.05$, paired t -test, see Supplementary Information). In all five rats, the firing patterns in the training square were reproduced in the test square (Fig. 4, square trials), while the circle-to-circle transfer results were more mixed (with some cells in rats 2 and 3 ceasing firing in the morph circle, see Fig. 4 cells 5, 6, 8, and a few firing in unrelated locations, see Fig. 4 cell 9). In summary, the five rats showed considerable tendency to generalize their geometrically tuned responses across perceptually different environments.

That environmental experience leads CA1 place cells to discriminate between environments on the basis of geometrical features, in the absence of differential reinforcement, and to maintain this discrimination for a long time, is consistent with the cognitive map theory of hippocampal function². (We note that rapid learning also plays an important role in this theory.) Further, the learning site(s) are most probably hippocampal, because cells in the superficial layers of entorhinal cortex, which comprise the dominant projection to the hippocampus, show inter-shape pattern similarity in exactly those circumstances in which hippocampal cells show divergence²³. We have captured hippocampal firing patterns throughout the animals' entire experience in two geometrically distinct enclosures, from initial entry, through to discrimination, transfer and delay stages. Our results identify a potential neural basis for hippocampal long-term memory, and may provide a model by which to test deficits in neuronal encoding, storage and retrieval, such as those induced by Alzheimer's disease. □

Methods

Twelve Lister Hooded rats maintained on a 12:12 hour light:dark schedule, weighing 280–425g at time of surgery, were used as subjects: 7 in the preliminary experiment, 5 in the main experiment (numbered 1 to 5 in the text). Methods were similar in both experiments. We describe the main experiment, pointing out differences where appropriate. See Supplementary Information for more details.

Testing procedures

Experiments were conducted in a black-curtained, circular testing arena 2.3 m in diameter. Rats were kept on a holding platform outside the arena before and after every trial. Two identical circular-walled (78 cm diameter), and two identical square-walled (59 cm sides), smooth, light-grey wooden boxes (all 50 cm high) were used alternately, placed on a platform 27 cm above the floor. This platform was cleaned before every trial. The centre of

each box always had the same location in the arena. A white cue card (58 cm wide, 50 cm high) placed at the north portion of the internal wall, and standardized procedures for placing the rat into the box, provided directional constancy. During trials the rat searched for grains of sweetened rice randomly thrown into the box about every 30 seconds. In the time-series component of the experiment, rats had alternating trials, circle first, in each shape for consecutive days. On days 1 and 2, rats had four trials, from day 3 onwards, six trials per day, and each rat had four trials for delay testing later on. Transfer trials used a 'morph' box made of 32 pieces of rectangular cross-section plastic tubing, 50 cm high, held together by packing tape and covered by masking tape on the inner surface. It could be configured as a 78-cm-diameter circle or a 62-cm-sided square. Rats 4 and 5, and the 7 rats of the preliminary experiment, were trained in the morph square and morph circle. For these animals an external cue card (102 cm high, 77 cm wide) suspended 25 cm inside the black-curtain rail provided directional constancy.

Monte Carlo simulation

Field peaks in the square or transformed circle are simulated as points drawn from a uniform distribution within a square box of side L . Points are binned, like the data, using 25×25 bins. The distances between one million pairs of points were found. The mean separation was $0.522L$ (s.d. $\pm 0.249L$), 5% had a separation $\leq 0.127L$, and 8.3% had separation $\leq 0.17L$.

Cell recording and data analysis

Methods of extracellular tetrode recording and firing-rate map construction were as previously reported^{21,30}. Rats in the main experiment were first run in the enclosures 3 to 8 weeks after surgical implantation of microdrives, once recording stability had been assessed on the holding platform outside the curtained testing arena. Place units were isolated offline and firing-rate maps were constructed by dividing the number of spikes by the rat's dwell time in a given location bin using customized software (Axona). The circle was transformed into a square using a standard algorithm (see Supplementary Information). Data from the square, circle, and transformed-circle were divided into 2.4-cm-sided square bins. The firing rate in each bin was smoothed using a boxcar average over the 5×5 surrounding bins centred on that bin. The five colours of firing-rate maps are each autoscaled to represent 20% of the peak rate (red to dark blue). Unvisited bins are shown in white.

Firing-rate maps from all cells and trials were inspected. We found no evidence of episode-specific, box-specific cell firing or any privileged time points at which remapping occurred. A few cells firing in only one trial per day were excluded as unreliable. In the main experiment, 30 cells from rat 1 (mean 5.0 per day), 264 cells from rat 2 (mean 12.6 per day), and 205 cells from rat 3 (mean 9.8 per day) contributed to the time-series data, from day 1 to 'dayend' (day 6 for rat 1, day 21 for rats 2 and 3). In all three rats, we recorded more cells later in the time series than early on.

Definitions of terms

The peak firing rate and its (x, y) location in the square or transformed-circle firing-rate map for each cell in each trial were used in the quantitative measures defined below for each cell on each day (detailed description in Supplementary Information). Rate divergence (RD): the ratio of the average absolute difference in peak firing rate between trials in different shapes over the average absolute difference in peak firing rate between trials in the same shape. Peak divergence (PD): ratio of the average distance between peak locations in different shapes over the average distance between peak locations in trials in the same shape. Cells with fields in both shapes of enclosure are classified as 'homotopic' according to the distance between the mean location of peak firing in the square and the mean location of peak firing in the transformed circle. A cell is 'homotopic' if this distance is less than expected at $P < 0.05$ for random fields, that is, $0.127L$ in the Monte Carlo simulation. We combined data from more than one animal to calculate field divergence (FD): the number of non-homotopic cells as a percentage of the total number of cells with fields in enclosures of both shape. Numbers quoted for 'similar'-looking fields correspond to fields with peak locations within $0.17L$ ($P < 0.083$ in the Monte Carlo simulations),

that is, it is a less strict criterion than homotopic. For analysis of the transfer test data, see Supplementary Information.

Received 8 August 2001; accepted 7 January 2002.

- Scoville, W. B. & Milner, B. Loss of recent memory after bilateral hippocampal lesions. *J. Neurol. Neurosurg. Psychiatr.* **20**, 11–21 (1957).
- O'Keefe, J. & Nadel, L. *The Hippocampus as a Cognitive Map* (Clarendon, Oxford, 1978).
- Squire, L. R. *Memory and Brain* (Oxford Univ. Press, New York, 1987).
- Cohen, N. J. & Eichenbaum, H. *Memory, Amnesia and the Hippocampal System* (MIT Press, Cambridge, MA, 1993).
- Bostock, E., Muller, R. U. & Kubie, J. L. Experience-dependent modifications of hippocampal place cell firing. *Hippocampus* **1**, 193–205 (1991).
- Kentros, C. *et al.* Abolition of long-term stability of new hippocampal place cell maps by NMDA receptor blockade. *Science* **280**, 2121–2126 (1998).
- Muller, R. U. & Kubie, J. L. The effects of changes in the environment on the spatial firing of hippocampal complex-spike cells. *J. Neurosci.* **7**, 1951–1968 (1987).
- Wilson, M. A. & McNaughton, B. L. Dynamics of the hippocampal ensemble code for space. *Science* **261**, 1055–1058 (1993).
- O'Keefe, J. & Speakman, A. Single unit activity in the rat hippocampus during a spatial memory task. *Exp. Brain Res.* **68**, 1–27 (1987).
- Mehta, M. R., Barnes, C. A. & McNaughton, B. L. Experience-dependent, asymmetric expansion of hippocampal place fields. *Proc. Natl Acad. Sci. USA* **94**, 8918–8921 (1997).
- Mehta, M. R., Quirk, M. C. & Wilson, M. A. Experience-dependent asymmetric fields. *Neuron* **25**, 707–715 (2000).
- Jeffery, K. J. in *Neuronal Mechanisms of Memory Formation* (ed. Holscher, C.) 100–121 (Cambridge Univ. Press, Cambridge, 2000).
- Merzenich, M. M. *et al.* Topographic reorganization of somatosensory cortical areas 3b and 1 in adult monkeys following restricted deafferentation. *Neuroscience* **8**, 33–55 (1983).
- Clark, S. A., Allard, T., Jenkins, W. J. & Merzenich, M. M. Receptive fields in the body-surface map in adult cortex defined by temporally correlated inputs. *Nature* **332**, 444–445 (1988).
- Wang, X., Merzenich, M. M., Sameshina, K. & Jenkins, W. M. Remodelling of hand representation in adult cortex determined by timing of tactile stimulation. *Nature* **378**, 71–75 (1995).
- Weinberger, N. M., Javid, R. & Lapan, B. Long-term retention of learning-induced receptive-field plasticity in the auditory cortex. *Proc. Natl Acad. Sci. USA* **90**, 2394–2398 (1993).
- O'Keefe, J., Nadel, L., Keightley, S. & Kill, D. Fornix lesions selectively abolish place learning in the rat. *Exp. Neurol.* **48**, 152–166 (1975).
- Morris, R. G., Garrud, P., Rawlins, J. N. & O'Keefe, J. Place navigation impaired in rats with hippocampal lesions. *Nature* **297**, 681–683 (1982).
- Sutherland, R. J., Kolb, B. & Whishaw, I. Spatial mapping: definitive disruption by hippocampal or medial frontal cortical damage in the rat. *Neurosci. Lett.* **31**, 271–276 (1982).
- Barnes, C. A. Spatial learning and memory processes: the search for their neurobiological mechanisms in the rat. *Trends Neurosci.* **11**, 163–169 (1988).
- O'Keefe, J. & Burgess, N. Geometric determinants of the place fields of hippocampal neurons. *Nature* **381**, 425–428 (1996).
- Hartley, T., Burgess, N., Lever, C., Cacucci, F. & O'Keefe, J. Modeling place fields in terms of the cortical inputs to the hippocampus. *Hippocampus* **10**, 369–379 (2000).
- Quirk, G. J., Muller, R. U., Kubie, J. L. & Ranck, J. B. The positional firing properties of medial entorhinal neurons: description and comparison with hippocampal place cells. *J. Neurosci.* **12**, 1945–1963 (1992).
- Sharp, P. E. Subicular cells generate similar spatial firing patterns in two geometrically and visually distinctive environments: comparison with hippocampal place cells. *Behav. Brain Res.* **85**, 71–92 (1997).
- Wood, E. R., Dudchenko, P. A. & Eichenbaum, H. The global record of memory in hippocampal neuronal activity. *Nature* **397**, 613–616 (1999).
- Frank, L. M., Brown, E. M. & Wilson, M. Trajectory encoding in the hippocampus and entorhinal cortex. *Neuron* **27**, 169–178 (2000).
- Tanila, H., Shapiro, M., Gallagher, M. & Eichenbaum, H. Brain aging: Changes in the nature of information coding by the hippocampus. *J. Neurosci.* **17**, 5155–5166 (1997).
- Hollup, S. A., Molden, S., Donnett, J. G., Moser, M. B. & Moser, E. I. Accumulation of hippocampal place fields at the goal location in an annular watermaze task. *J. Neurosci.* **21**, 1635–1644 (2001).
- O'Keefe, J. & Recce, M. L. Phase relationship between hippocampal place units and the EEG theta rhythm. *Hippocampus* **3**, 317–330 (1993).
- Jeffery, K. J. & O'Keefe, J. Learned interaction of visual and idiothetic cues in the control of place field orientation. *Exp. Brain Res.* **127**, 151–161 (1999).

Supplementary Information accompanies the paper on *Nature's* website (<http://www.nature.com>).

Acknowledgements

We thank J. Huxter, T. Hartley and K. Jeffery for discussion, S. Burton for reviewing the manuscript, C. Parker for technical assistance, and J. Donnett for the recording system. This work was supported by a Medical Research Council (UK) programme grant to J.O.K., MRC Advanced Studentships to C.L. and T.W., a Wellcome Trust Studentship to F.C., and Royal Society and MRC Senior Fellowships to N.B.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.L. (e-mail: colin@maze.ucl.ac.uk) or J.O.K. (e-mail: John@maze.ucl.ac.uk).

Balanced responsiveness to chemoattractants from adjacent zones determines B-cell position

Karin Reif[†], Eric H. Ekland[†], Lars Ohl[‡], Hideki Nakano[§], Martin Lipp^{||}, Reinhold Förster[‡] & Jason G. Cyster^{*}

^{*} Howard Hughes Medical Institute and Department of Microbiology and Immunology, University of California San Francisco, 513 Parnassus Avenue, San Francisco, California 94143-0414, USA

[‡] University Clinic for Surgery, Nikolaus-Fiebiger-Center, 91054 Erlangen, Germany

[§] Department of Immunology, Toho University School of Medicine, Ota-Ku, Tokyo 143-8540, Japan

^{||} Molecular Tumorigenetics and Immunogenetics, Max-Delbrück-Center for Molecular Medicine, 13092 Berlin, Germany

[†] These authors contributed equally to this work

B lymphocytes re-circulate between B-cell-rich compartments (follicles or B zones) in secondary lymphoid organs, surveying for antigen. After antigen binding, B cells move to the boundary of B and T zones to interact with T-helper cells^{1–3}. Despite the importance of B–T-cell interactions for the induction of antibody responses, the mechanism causing B-cell movement to the T zone has not been defined. Here we show that antigen-engaged B cells have increased expression of CCR7, the receptor for the T-zone chemokines^{4,5} CCL19 and CCL21, and that they exhibit increased responsiveness to both chemoattractants. In mice lacking lymphoid CCL19 and CCL21 chemokines, or with B cells that lack CCR7, antigen engagement fails to cause movement to the T zone. Using retroviral-mediated gene transfer we demonstrate that increased expression of CCR7 is sufficient to direct B cells to the T zone. Reciprocally, overexpression of CXCR5, the receptor for the B-zone chemokine CXCL13, is sufficient to overcome antigen-induced B-cell movement to the T zone. These findings define the mechanism of B-cell relocation in response to antigen, and establish that cell position *in vivo* can be determined by the balance of responsiveness to chemoattractants made in separate but adjacent zones.

Using B cells from mice carrying immunoglobulin transgenes specific for the antigen hen-egg lysozyme (Ig^{HEL})⁶ in a time course analysis, we found that movement of antigen-engaged B cells to the B-zone/T-zone (B/T) boundary is complete within 6 h of antigen exposure (Fig. 1a). Antigen-engaged B cells can remain at this location for at least two days², and interactions between T-helper cells and B cells are first observed in this zone³. Genetic studies have established that CCR7, and its ligands CCL19 and CCL21, are important for T-cell and dendritic cell localization in T-cell zones^{7,8}. To test whether movement of B cells to the T zone might occur owing to increased expression of CCR7, we tested for levels of CCL19-Fc (Fc, fragment crystallizable) binding to antigen-stimulated B cells (Fig. 1b). Six hours after exposure to HEL antigen *in vivo*, Ig^{HEL}-transgenic B cells showed a two–threefold increase in binding of CCL19-Fc, and little or no change in CXCR5 expression (Fig. 1b). CCL19-Fc binding to CCR7-deficient B cells was minimal, establishing that CCL19-Fc staining served as a faithful reporter of CCR7 levels (see Supplementary Information). Efforts to measure how the increase in CCR7 affected chemotactic responsiveness of the cells were confounded, as B cells that had bound large amounts of HEL antigen attached irreversibly to transwell filters (data not shown), consistent with reports that HEL binds strongly to many surfaces⁹.

We therefore developed a second approach to study antigen-engaged B-cell redistribution, incubating B cells from IgM^a con-

genic C57BL/6 (B6) mice *in vitro* with anti-IgM antibody as a surrogate antigen for 1 h, and then transferring the cells to IgM^b B6 mice. The anti-IgM treatment caused the cells to localize at the B/T boundary (Fig. 1c), and induced upregulation of CCL19-Fc binding, which was equivalent to the induction in HEL binding observed with Ig^{HEL}-transgenic B cells (Fig. 1b). *Ex vivo* chemotaxis assays with anti-IgM-stimulated cells revealed an increase in the CCL19 and CCL21 chemokine response profile of the cells, with an increase in the magnitude of the response particularly to low doses of chemokines (Fig. 1d). The basis for the greater augmentation of CCL19 compared with CCL21 responsiveness is unknown. There was no significant change in responsiveness to the B-zone chemo-

kine CXCL13 after B-cell antigen receptor (BCR) stimulation (Fig. 1d), consistent with the lack of change in CXCR5 expression (Fig. 1b).

Our *ex vivo* studies indicated that antigen-mediated relocation of B cells to the T zone requires B-cell expression of CCR7 and response to the T-zone chemokines CCL19 and CCL21. To test this *in vivo*, we transferred Ig^{HEL}-transgenic B cells to *plt/plt* mice that lack the *CCL19* gene^{10,11} and the gene for the lymphoid tissue form of *CCL21* (refs 7, 12). In the absence of antigen, Ig^{HEL}-transgenic B cells localized in follicles, consistent with previous findings that T-zone chemokines and CCR7 are not necessary for follicle formation. However, HEL antigen exposure failed to cause redistribution of

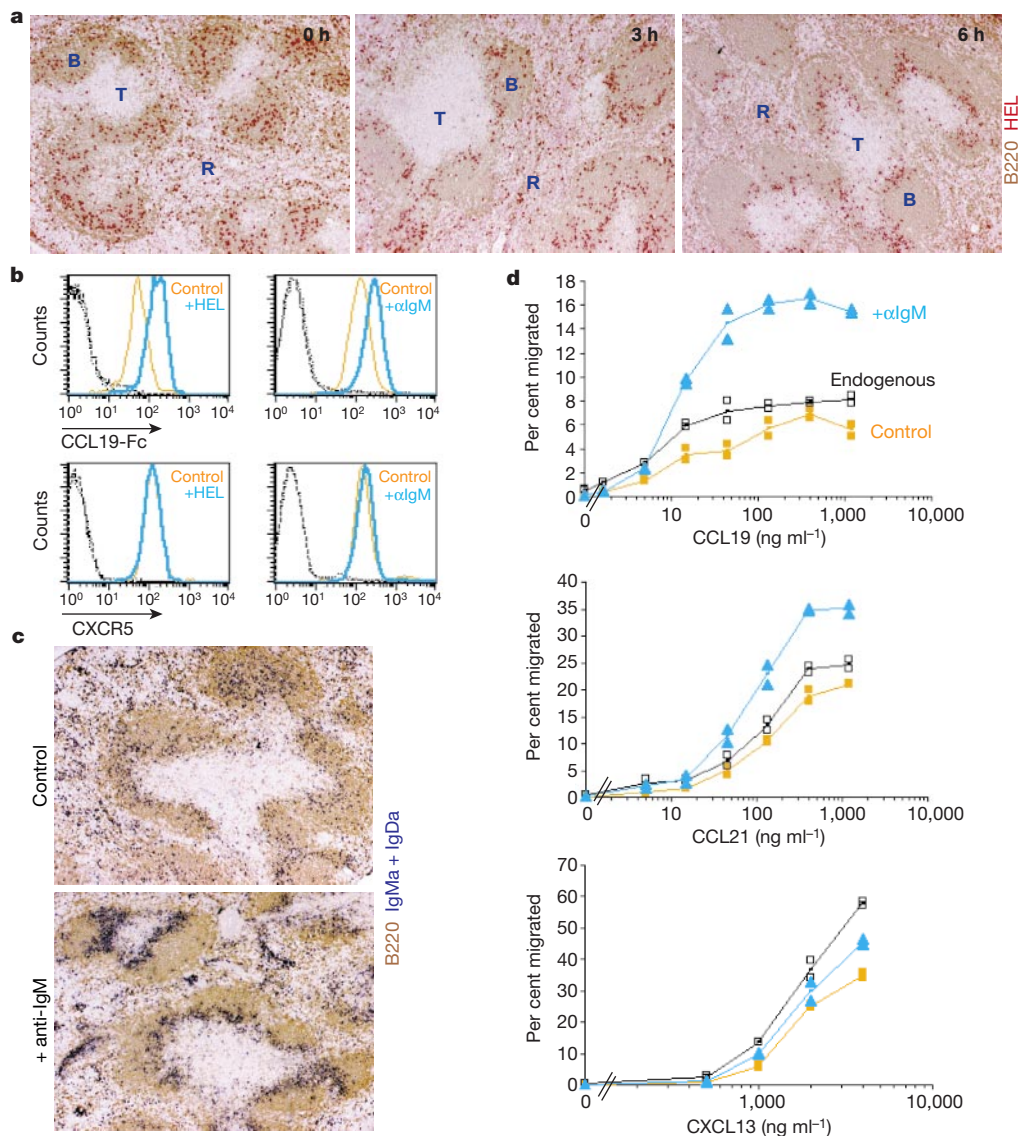


Figure 1 Antigen-engaged B cells upregulate CCR7, increase responsiveness to T-zone chemokines and migrate to the outer T zone. **a**, Mice that received Ig^{HEL}-transgenic B cells were injected with HEL, and the distribution in the spleen of HEL-binding cells (red) and total B cells (brown) was determined 0, 3 and 6 h later. B, B zone; T, T zone; R, red pulp. **b**, CCL19-Fc and CXCR5 staining of Ig^{HEL}-transgenic B cells isolated from mice 6 h after saline (control) or HEL injection (left panels), or IgM^a non-transgenic B cells isolated 5.5 h after transfer of cells that had been incubated *in vitro* in the absence (control) or presence of anti-IgM (α-IgM; right panels). Blue line, stimulated cells; orange line, control cells; dotted and dashed black lines, control or stimulated cells stained with human LFA3-Fc

(upper panels) or without primary antibody (lower panels) as controls. **c**, Distribution of IgM^a B cells incubated for 1 h without (control) or with anti-IgM at 5.5 h after transfer to IgM^b B6 recipient mice. Spleen sections were stained for IgM^a and IgD^a (blue stain) to detect transferred B cells, and for B220 (brown) to delineate B-cell follicles. **d**, *Ex vivo* chemotaxis assays of spleen cells from transfer recipient mice. The percentage of input B cells of the indicated type that migrated in response to chemokine is shown in duplicate. The response of stimulated cells to CCL19 and CCL21 was significantly greater than the response of control cells (CCL19, $P < 0.01$; CCL21, $P < 0.05$). Data in **a–d** are representative of more than 3 experiments of each type.

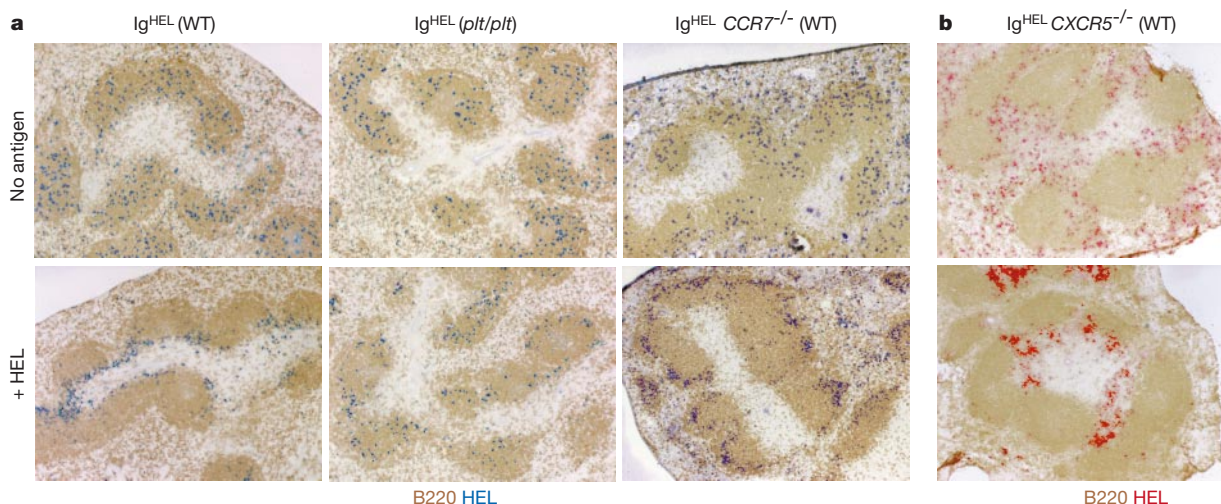


Figure 2 Localization of antigen-engaged B cells along the B/T boundary is dependent on both T- and B-zone chemokines. **a**, Spleen sections of wild-type (WT) or *plt/plt* mice that received wild-type or *CCR7*^{-/-} Ig^{HEL}-transgenic B cells, as indicated, 6 h after injection of saline (no antigen) or HEL antigen, stained to detect HEL-binding B cells (blue) and total

B cells (brown). **b**, Spleen sections of wild-type mice that received *CXCR5*^{-/-} Ig^{HEL}-transgenic B cells, prepared and stained as in **a**, except HEL-binding cells are red. Data are representative of more than 10 experiments for *plt/plt* recipients, three for *CCR7*^{-/-} cells, and four for *CXCR5*^{-/-} cells.

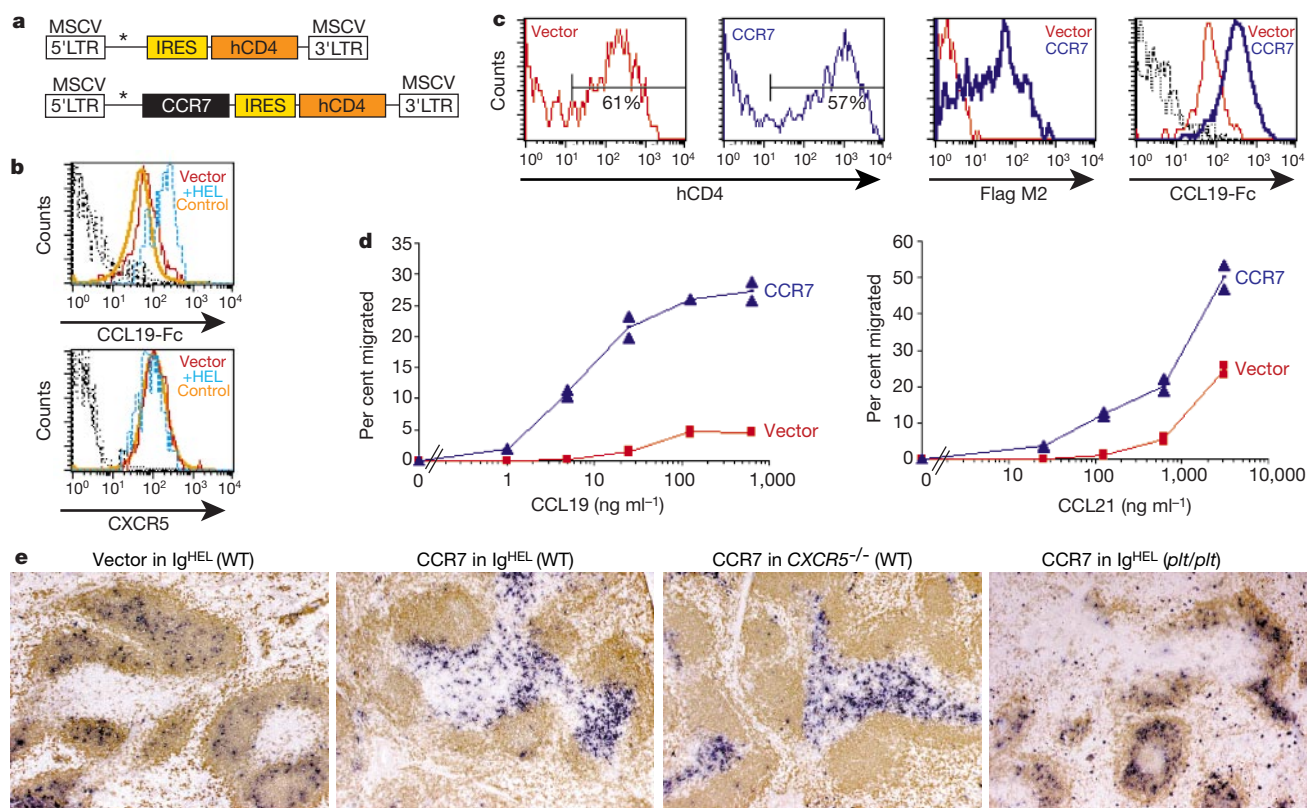


Figure 3 Increased expression of CCR7 is sufficient to mediate B-cell redistribution to the T zone. **a**, Diagram of the MSCV.2 retroviral vector and Flag-CCR7-containing vector. MSCV, murine stem cell virus; LTR, long-term repeat; asterisk, packaging signal; IRES, internal ribosomal entry site; hCD4, cytoplasmic-domain-truncated human CD4. **b**, *Ex vivo* flow cytometric analysis of cells from recipient mice of CFSE-labelled Ig^{HEL}-transgenic B cells. CD40-mediated activation and retroviral infection do not alter CCL19-Fc staining or CXCR5 levels. Red line, vector-infected cells from mice lacking HEL; dashed blue line, vector-infected cells from mice injected with HEL; orange line, endogenous B cells; dotted and dashed black lines, staining controls for above cells (see Fig. 1b). **c**, *Ex vivo* flow cytometric analysis of cells from recipient mice of CFSE-labelled Ig^{HEL}-transgenic B cells transduced with empty (vector) or Flag-CCR7-containing (CCR7) retroviral vectors. The fraction of transferred B cells expressing human CD4 (two left panels), and Flag-epitope

(Flag M2) and CCL19-Fc staining of the human CD4-positive cells (two right panels), is shown. Blue line, CCR7-infected cells; red line, vector-infected cells; dotted and dashed black lines, control-infected cells or CCR7-infected cells, respectively, stained with human LFA3-Fc as staining control. **d**, *Ex vivo* chemotaxis assays with spleen cells of mice that had received, one day before, Ig^{HEL}-transgenic B cells transduced with the indicated retrovirus. The percentage of input B220⁺ human CD4⁺ control- (vector) or CCR7-infected cells that migrated to the indicated chemokine, is shown. **e**, Splenic distribution of transferred Ig^{HEL}-transgenic wild-type (WT) or *CXCR5*^{-/-} B cells infected with empty (vector) or CCR7-containing virus, stained to detect transduced cells (human CD4, blue) in relation to B-cell follicles (B220, brown), one day after transfer to wild-type or *plt/plt* recipient mice, as indicated. Data in **b–e** are representative of at least three separate experiments.

cells to the outer T zone (Fig. 2a), indicating a role for CCL19 and/or CCL21. A concern with this approach was that the effect could be indirect, owing to alterations in the T zone that were downstream of T-zone chemokine deficiency. To test for an intrinsic requirement for CCR7 in antigen-engaged B cells, we introduced the Ig^{HEL} transgenes onto the CCR7 knockout background⁸, and performed a similar transfer experiment. HEL-binding CCR7-deficient B cells failed to localize in the outer T zone (Fig. 2a); notably, the antigen-engaged CCR7-deficient B cells often moved in the opposite direction, lodged in the pole of the follicle distal to the T-zone.

The flow cytometric and *ex vivo* chemotaxis analyses of antigen-stimulated B cells established that the cells retain CXCR5 expression and responsiveness to CXCL13 (Fig. 1). To examine whether CXCL13 responsiveness influenced the distribution of antigen-binding B cells, CXCR5^{-/-} Ig^{HEL}-transgenic B cells were transferred to mice in the absence or presence of the HEL antigen. As described previously^{13,14}, CXCR5-deficient cells did not enter follicles, and many of the cells were found in the red pulp, with some cells localized at entry points (bridging zones) to the T-cell areas (Fig. 2b). Antigen binding caused a marked increase in the accumulation of cells at the edge of the T zone, confirming that CXCR5 is not required for cells to move to this compartment in response to antigen (Fig. 2b). However, the CXCR5-deficient cells failed to distribute uniformly along the B/T boundary, and were found more often in clusters near bridging zones (Fig. 2b). Therefore, CXCL13 influences the localization of antigen-engaged B cells,

helping to distribute the cells along the B/T boundary.

The above findings suggested that the increased surface expression of CCR7 induced by BCR signalling might be the principal determinant causing B-cell relocalization to the T zone. To test this model, we used retroviral-mediated gene transfer^{15–17} to selectively alter chemokine receptor levels in B cells, followed by adoptive transfer into mice (Fig. 3). To activate the B cells for retroviral infection, we incubated Ig^{HEL}-transgenic cells with mitogenic doses of anti-CD40 antibody. In contrast to BCR triggering, cells activated by CD40 remain competent for migration into follicles, and do not undergo major changes in expression of CXCR5 or CCR7 (Fig. 3b). As expected, transduction with Flag-tagged retrovirus containing CCR7 led to co-expression of the marker protein—cytoplasmic-domain-truncated human CD4 (Fig. 3a)—and Flag-CCR7 on the surface of most of the transferred B cells (Fig. 3c), and increased CCL19-Fc binding by these cells (Fig. 3c). *Ex vivo* chemotaxis analysis of the transduced cells revealed that the higher CCR7 expression conferred an increase in sensitivity to CCL19 and CCL21 (Fig. 3d). This outcome was similar to the effect of BCR stimulation (Fig. 1d), although of a greater magnitude, consistent with the higher expression of CCR7 on retrovirally transduced cells (four–fivefold increase) compared with BCR-stimulated cells (two–threefold increase). Using human CD4 staining as a surrogate marker for tracking B cells that over-express CCR7 *in situ*, we found that the cells distributed within the T zone and failed to localize within B-cell follicles (Fig. 3e). By

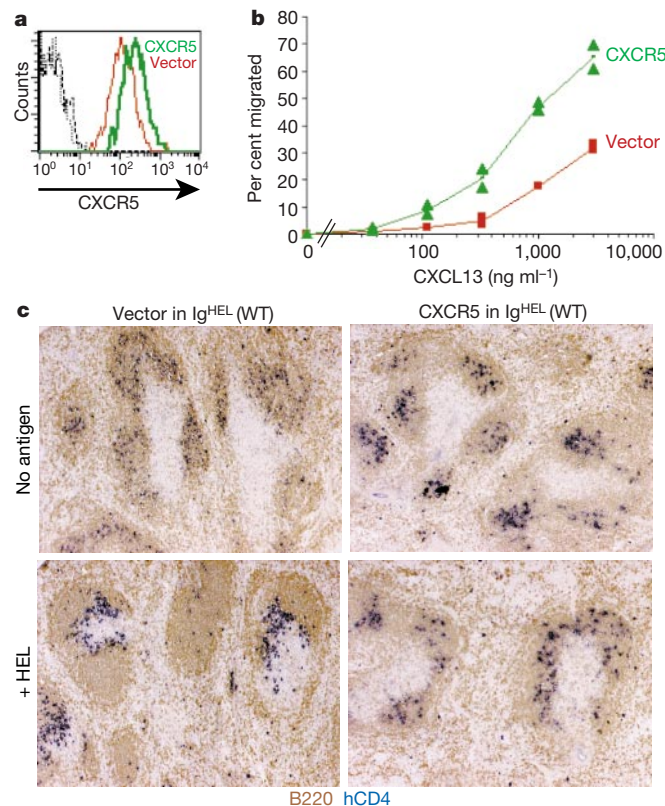


Figure 4 Overexpression of CXCR5 is sufficient to override antigen-induced B-cell movement to the T-cell zone. **a**, *Ex vivo* flow cytometric analysis of spleen cells from recipient mice of CFSE-labelled Ig^{HEL}-transgenic B cells transduced with empty or CXCR5-containing retroviral vector. Green line, CXCR5-infected cells; red line, vector-infected cells; dotted and dashed black lines, control-infected cells or CXCR5-infected cells, respectively, stained without primary antibody as staining control. **b**, *Ex vivo* chemotaxis assays with spleen cells of mice that, one day earlier, had received Ig^{HEL}-transgenic B

cells transduced with the indicated retrovirus. Assays were as in Fig. 3. **c**, Distribution in spleen of transferred Ig^{HEL}-transgenic wild-type (WT) cells that had been infected with empty (vector) or CXCR5-containing virus, and transferred to wild-type mice that, one day later, received an injection of saline or HEL antigen, as indicated. Tissue was isolated 7 h later and stained to detect human CD4 (blue) and B220 (brown). Data in **a–c** are representative of three separate experiments.

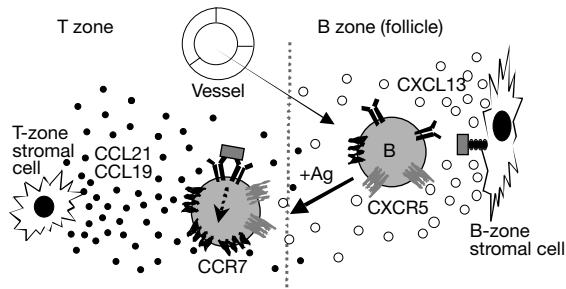


Figure 5 Schematic representation of the mechanism for B-cell movement to the B/T boundary after BCR engagement. B cells enter spleen and lymph nodes through blood vessels that are mostly located outside B-cell zones. Naive B cells express high levels of CXCR5 and low levels of CCR7, and their responsiveness to the CXCR5 ligand, CXCL13, dominates, causing them to move into follicles. When antigen is encountered, BCR signalling leads to increased surface CCR7 and increased responsiveness to the CCR7 ligands, CCL19 and CCL21. The shifted balance of chemokine responsiveness causes the antigen-engaged cell to move to the T zone. Open circles, CXCL13; filled circles, CCL19 or CCL21; rectangular box, antigen; antigen-retaining structure on stromal cell, complement receptor; Y-shaped structure on B cell, BCR.

contrast, cells transduced with the empty vector localized within follicles in a similar way to non-transduced cells (Fig. 3e). The CCR7-transduced B cells accumulated frequently along the B/T boundary in greater numbers than in other parts of the T zone (Fig. 3e). This effect was not seen when the CCR7-transduced cells lacked CXCR5 (Fig. 3e), establishing that accumulation near the boundary was dependent on responsiveness to CXCL13. As expected, movement of CCR7-transduced B cells to the T zone was dependent on CCL19 and/or CCL21, as the cells failed to lodge in this zone in *plt/plt* mice (Fig. 3e).

Antigen-engaged B cells are responsive to both T- and B-zone chemokines, and the findings shown in Figs 2 and 3 indicate that responsiveness to both types of chemokine contribute towards positioning the cells. To further explore the extent to which antigen-engaged B-cell positioning reflects the outcome of a balance of chemokine responsiveness, we tested whether overexpression of CXCR5 by retroviral transduction (Fig. 4a) is sufficient to overcome the effect of antigen engagement on B-cell distribution. Adoptively transferred cells overexpressing CXCR5 showed a marked increase in responsiveness to CXCL13 in *ex vivo* chemotaxis assays (Fig. 4b). BCR engagement failed to promote follicular exclusion of cells overexpressing CXCR5, and the cells were found in follicles, within the zone of CXCL13 expression (Fig. 4c and data not shown). Cells transduced with empty vector were excluded from follicles in a manner similar to non-transduced cells (Fig. 4c).

These findings establish that the positioning of B cells in lymphoid tissues *in vivo* is determined by the relative responsiveness of the cells to chemokines made in separate but adjacent zones. In naive B cells, the balance of responsiveness is in favour of CXCL13, and the cells migrate into follicles. Upon antigen encounter and BCR stimulation, a change in the balance occurs that favours the response to T-zone chemokines, and this change is sufficient to redistribute antigen-engaged B cells to the T zone (Fig. 5). Our observations suggest that B cells integrate signals from two chemokine receptors simultaneously, consistent with *in vitro* studies showing that neutrophils can sense and respond to two sources of chemoattractant¹⁸. Antigen-stimulated B cells respond less strongly than T cells and dendritic cells to CCL19 and CCL21 (refs 19, 20), and we propose that this contributes to their exclusion from the central T zone. Reduced CCR7 expression in T-helper subtype 2 cell lines has been associated with the failure of these cells to reach the central T zone¹⁷. The possible contribution of additional factors, such as adhesion molecules, to placement of antigen-engaged B cells

at the B/T boundary are not ruled out by our study. In addition, although we have focused on the behaviour of B cells in response to foreign antigen, autoreactive B cells that are chronically engaged by auto-antigen also lodge at the B/T boundary before their elimination^{21,22}. It is probable that a similar mechanism operates to regulate the localization of auto-antigen-binding cells. We suggest that the positioning of many cell types is determined by balanced responsiveness to chemoattractants emanating from adjacent zones. This mechanism may be used in the positioning of other types of leukocytes, such as activated T cells and subsets of dendritic cells^{3,17,23}, in the localization of growing axons, and in the patterning of organs during development²⁴. □

Methods

Mice, adoptive cell transfer and stimulation procedures

Ig^{HEL}-transgenic mice were of the MD4 line⁶, *plt/plt* mice²⁵ on a BALB/c background were crossed to C57BL/6 (B6) mice for six generations; CXCR5^{-/-} mice¹³ were eighth-generation backcrossed to B6 mice, and CCR7^{-/-} mice⁸ were on a mixed B6/129 background. Igha (Ig heavy chain of a) Thy1a Gpi1a B6 mice (termed IgM^a congenic B6) were from Jackson Laboratory. B6 (IgM^b) recipient mice were from Charles River or a colony maintained at University of California San Francisco. To provide CCR7^{-/-} cells for adoptive transfer, lethally irradiated B6 mice were reconstituted for five weeks with CCR7^{-/-} Ig^{HEL} bone marrow, as described². Adoptive transfers were performed using 1.5–3 × 10⁷ spleen cells per recipient mouse². In some experiments, mice were injected intraperitoneally 12–18 h after the cell transfer with 0.8–1 mg HEL (Sigma) in phosphate-buffered saline (PBS) or with PBS alone. For IgM stimulation, splenocytes were cultured for 1 h at 37 °C in RPMI 1640 supplemented with 3% FBS, 50 µg ml⁻¹ streptomycin, 50 U ml⁻¹ penicillin, 10 mM HEPES with or without 10 µg ml⁻¹ goat anti-mouse IgM F(ab')₂ polyclonal antibodies (Jackson ImmunoResearch), washed and then transferred. Spleens were collected 5.5–6 h after stimulation or at the time indicated in the results, and cut into fragments with 1 out of 3 being used to generate cell suspensions for *ex vivo* flow cytometry and chemotaxis assays, and 2 out of 3 being frozen in OCT compound (Sakura).

Retroviral transduction of B cells

Flag-tagged CCR7 (ref. 19) or CXCR5 (ref. 26) were cloned into the MSCV2.2 retroviral vector containing cytoplasmic-domain-truncated human CD4 as an expression marker downstream of the internal ribosomal entry site (IRES)¹⁷. Retrovirus containing supernatant was generated using the Bosc23 packaging cell line together with pCL-Eco packaging vector²⁷, as described²⁸. For B-cell activation, Ig^{HEL}-transgenic or CXCR5^{-/-} Ig^{HEL}-transgenic mice were injected intraperitoneally with 4–6 mg HEL, splenocytes were collected after 6–8 h, cultured in RPMI 1640 supplemented with 10% FBS, 2 mM L-glutamine, 50 µg ml⁻¹ streptomycin, 50 U ml⁻¹ penicillin, 10 mM HEPES, and 50 µM 2-ME in the presence of 10 µg ml⁻¹ anti-CD40 monoclonal antibody (clone FGK45, a gift from A. Rolink), and spin-infected 24 h later with retroviral supernatants and 4 µg ml⁻¹ polybrene (Sigma). Two days after infection, cells were collected, in some experiments labelled with 2 µM CFSE, and adoptively transferred into mice. In all experiments, between 45% and 75% of B cells were retrovirally transduced as assessed by flow cytometric analysis for human CD4 expression. T cells were not activated by the antigen and anti-CD40 treatment, and were not detectably infected by the retroviruses.

Chemotaxis assays, flow cytometry and immunohistochemistry

For chemotaxis assays, cells transmigrated across 5-µm transwell filters (Corning Costar Corp) for over 3 h in response to a stimulus in the bottom chamber, and were enumerated by flow cytometry¹⁹. Chemokines were provided by R&D Systems. Transwell assays were performed in duplicate for each chemokine concentration, and were repeated using cells from a minimum of three different animals of each type. We carried out statistical analyses using the paired *t*-test. The CCL19-Fc recombinant protein was a fusion of mouse CCL19 and the human IgG1 constant region²⁹. For control staining, we used human LEA3-Fc (a gift from J. Brown). Antibodies used were anti-CXCR5 (refs 13, 30), anti-human CD4-phycoerythrin (BD Pharmingen), and biotinylated Flag M2 (Sigma). We analysed labelled cells on a FACSCaliber (Becton Dickinson). See Supplementary Information for CCL19-Fc background staining on CCR7^{-/-} cells, and anti-CXCR5 background staining on CXCR5^{-/-} B cells. Immunohistochemical analysis was performed as described¹⁴, with the exception that HEL-binding cells were detected by incubation with 1–5 µg ml⁻¹ HEL antigen followed by biotin-conjugated rabbit anti-HEL (Rockland). Human CD4 was detected with biotin-conjugated mouse anti-human CD4 (BD Pharmingen). Biotinylated antibodies were visualized with alkaline phosphatase-conjugated streptavidin (Jackson ImmunoResearch) and Fast Red TR or Fast Blue BB salt (Sigma) as a substrate. All images of spleen sections were obtained at × 5 objective magnification, and each image shown is representative of all (10–20) white pulp cords observed in three transverse sections of each spleen.

Received 8 November 2001; accepted 10 January 2002.

1. Liu, Y.-J., Oldfield, S. & MacLennan, I. C. M. Memory B cells in T cell-dependent antibody responses colonize the splenic marginal zones. *Eur. J. Immunol.* **18**, 355–362 (1988).
2. Cyster, J. G. & Goodnow, C. C. Antigen-induced exclusion from follicles and anergy are separate and complementary processes that influence peripheral B cell fate. *Immunity* **3**, 691–701 (1995).

3. Garside, P. *et al.* Visualization of specific B and T lymphocyte interactions in the lymph node. *Science* **281**, 96–99 (1998).
4. Cyster, J. G. Chemokines and cell migration in secondary lymphoid organs. *Science* **286**, 2098–2102 (1999).
5. Zlotnik, A. & Yoshie, O. Chemokines: a new classification system and their role in immunity. *Immunity* **12**, 121–127 (2000).
6. Goodnow, C. C. *et al.* Altered immunoglobulin expression and functional silencing of self-reactive B lymphocytes in transgenic mice. *Nature* **334**, 676–682 (1988).
7. Gunn, M. D. *et al.* Mice lacking expression of secondary lymphoid organ chemokine have defects in lymphocyte homing and dendritic cell localization. *J. Exp. Med.* **189**, 451–460 (1999).
8. Förster, R. *et al.* CCR7 coordinates the primary immune response by establishing functional microenvironments in secondary lymphoid organs. *Cell* **99**, 23–33 (1999).
9. Basten, A. *et al.* Self tolerance in the B cell repertoire. *Immunol. Rev.* **122**, 5–19 (1991).
10. Luther, S. A., Tang, H. L., Hyman, P. L., Farr, A. G. & Cyster, J. G. Coexpression of the chemokines ELC and SLC by T zone stromal cells and deletion of the ELC gene in the *plt/plt* mouse. *Proc. Natl Acad. Sci. USA* **97**, 12694–12699 (2000).
11. Nakano, H. & Gunn, M. D. Gene duplications at the chemokine locus on mouse chromosome 4: multiple strain-specific haplotypes and the deletion of secondary lymphoid-organ chemokine and EBI-1 ligand chemokine genes in the *plt* mutation. *J. Immunol.* **166**, 361–369 (2001).
12. Vassileva, G. *et al.* The reduced expression of 6CKine in the *plt* mouse results from the deletion of one of two 6CKine genes. *J. Exp. Med.* **190**, 1183–1188 (1999).
13. Förster, R. *et al.* A putative chemokine receptor, BLR1, directs B cell migration to defined lymphoid organs and specific anatomic compartments of the spleen. *Cell* **87**, 1037–1047 (1996).
14. Ansel, K. M. *et al.* A chemokine driven positive feedback loop organizes lymphoid follicles. *Nature* **406**, 309–314 (2000).
15. Ouyang, W. *et al.* Inhibition of Th1 development mediated by GATA-3 through an IL-4-independent mechanism. *Immunity* **9**, 745–755 (1998).
16. Van Parijs, L., Refaelli, Y., Abbas, A. K. & Baltimore, D. Autoimmunity as a consequence of retrovirus-mediated expression of C-FLIP in lymphocytes. *Immunity* **11**, 763–770 (1999).
17. Randolph, D. A., Huang, G., Carruthers, C. J. L., Bromley, L. E. & Chaplin, D. D. Differential expression of CCR7 on Th1 and Th2 cells: Importance in control of T cell localization and delivery of help for antibody responses *in vivo*. *Science* **286**, 2159–2162 (1999).
18. Foxman, E. F., Kunkel, E. J. & Butcher, E. C. Integrating conflicting chemotactic signals. The role of memory in leukocyte navigation. *J. Cell Biol.* **147**, 577–588 (1999).
19. Ngo, V. N., Tang, H. L. & Cyster, J. G. Epstein–Barr virus-induced molecule 1 ligand chemokine is expressed by dendritic cells in lymphoid tissues and strongly attracts naive T cells and activated B cells. *J. Exp. Med.* **188**, 181–191 (1998).
20. Kellermann, S. A., Hudak, S., Oldham, E. R., Liu, Y. J. & McEvoy, L. M. The CC chemokine receptor-7 ligands 6CKine and macrophage inflammatory protein-3 beta are potent chemoattractants for *in vitro*- and *in vivo*-derived dendritic cells. *J. Immunol.* **162**, 3859–3864 (1999).
21. Cyster, J. G., Hartley, S. B. & Goodnow, C. C. Competition for follicular niches excludes self-reactive cells from the recirculating B-cell repertoire. *Nature* **371**, 389–395 (1994).
22. Mandik-Nayak, L., Bui, A., Noorchashm, H., Eaton, A. & Erikson, J. Regulation of anti-double-stranded DNA B cells in nonautoimmune mice: localization to the T–B interface of the splenic follicle. *J. Exp. Med.* **186**, 1257–1267 (1997).
23. Iwasaki, A. & Kelsall, B. L. Localization of distinct Peyer's patch dendritic cell subsets and their recruitment by chemokines macrophage inflammatory protein (MIP)-3α, MIP-3β, and secondary lymphoid organ chemokine. *J. Exp. Med.* **191**, 1381–1394 (2000).
24. Belloch, R., Newman, C. & Kimble, J. Control of cell migration during *Caenorhabditis elegans* development. *Curr. Opin. Cell Biol.* **11**, 608–613 (1999).
25. Nakano, H. *et al.* Genetic defect in T lymphocyte-specific homing into peripheral lymph nodes. *Eur. J. Immunol.* **27**, 215–221 (1997).
26. Gunn, M. D. *et al.* A B-cell-homing chemokine made in lymphoid follicles activates Burkitt's lymphoma receptor-1. *Nature* **391**, 799–803 (1998).
27. Naviaux, R. K., Costanzi, E., Haas, M. & Verma, I. M. The pCL vector system: rapid production of helper-free, high-titer, recombinant retroviruses. *J. Virol.* **70**, 5701–5705 (1996).
28. Pear, W. S., Nolan, G. P., Scott, M. L. & Baltimore, D. Production of high-titer helper-free retroviruses by transient transfection. *Proc. Natl Acad. Sci. USA* **90**, 8392–8396 (1993).
29. Hargreaves, D. C. *et al.* A coordinated change in chemokine responsiveness guides plasma cell movements. *J. Exp. Med.* **194**, 45–56 (2001).
30. Schmidt, K. N., Hsu, C. W., Griffin, C. T., Goodnow, C. C. & Cyster, J. G. Spontaneous follicular exclusion of SHP1-deficient B cells is conditional on the presence of competitor wild-type B cells. *J. Exp. Med.* **187**, 929–937 (1998).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com>).

Acknowledgements

We thank K. Murphy and W. Sha for retroviral vectors; P. Andres, D. Bhattacharya, Y. Refaelli and W. Sha for advice; M. Matloubian, T. Okada, D. Stainier and A. Weiss for comments on the manuscript; and A. Bidgol for help with screening mice. K.R. was supported by a Human Frontier Science Program Long Term Fellowship and is currently supported by a Leukemia and Lymphoma Society Special Fellowship. E.H.E. is a Howard Hughes Medical Institute (HHMI) predoctoral fellow, and J.G.C. is a HHMI assistant investigator and a Packard Fellow. This work was supported in part by a grant from the National Institutes of Health.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.G.C. (e-mail: cyster@itsa.ucsf.edu).

Ubiquitination-dependent cofactor exchange on LIM homeodomain transcription factors

Heather P. Ostendorff[†], Reto I. Peirano[†], Marvin A. Peters^{*}, Anne Schlüter^{*}, Michael Bossenz^{*}, Martin Scheffner[‡] & Ingolf Bach^{*}

^{*} Zentrum für Molekulare Neurobiologie Hamburg (ZMNH), Universität Hamburg, Martinistrasse 85, 20251 Hamburg, Germany
[‡] Institut für Biochemie I, Medizinische Fakultät, Universität zu Köln, Joseph-Stelzmann-Strasse 52, 50931 Köln, Germany
[†] These authors contributed equally to this work

The interactions of distinct cofactor complexes with transcription factors are decisive determinants for the regulation of gene expression. Depending on the bound cofactor, transcription factors can have either repressing or transactivating activities¹. To allow a switch between these different states, regulated cofactor exchange has been proposed^{2,3}; however, little is known about the molecular mechanisms that are involved in this process. LIM homeodomain (LIM-HD) transcription factors associate with RLIM (RING finger LIM domain-binding protein) and with CLIM (cofactor of LIM-HD proteins; also known as NLI, Ldb and Chip) cofactors. The co-repressor RLIM inhibits the function of LIM-HD transcription factors, whereas interaction with CLIM proteins is important for the exertion of the biological activity conferred by LIM-HD transcription-factors^{4,5}. Here we identify RLIM as a ubiquitin protein ligase that is able to target CLIM cofactors for degradation through the 26S proteasome pathway. Furthermore, we demonstrate a ubiquitination-dependent association of RLIM with LIM-HD proteins in the presence of CLIM cofactors. Our data provide a mechanistic basis for cofactor exchange on DNA-bound transcription factors, and probably represent a general mechanism of transcriptional regulation.

We have used the regulation of LIM-HD transcription factors^{4,5} as a model system to investigate mechanisms of controlling transcription factor activity. Their interaction with CLIM cofactors is thought to be critical for the exertion of the transcriptional and biological activity of this class of developmental regulators^{6–11}. In contrast, LIM-HD protein activity is negatively regulated by nuclear LIM-only (LMO) proteins^{9,10,12}—which compete with LIM-HD proteins for CLIM cofactors—and by the RING H2 zinc finger protein RLIM. Indeed, overexpression of full-length RLIM or a dominant negative version of CLIM (DN-CLIM) in the developing chick wing results in very similar phenotypes, inhibiting the development of distal wing structures⁸.

The ubiquitin proteasome-dependent proteolytic pathway is involved in the regulation of critical developmental regulator proteins such as NF-κB, c-Jun, Tramtrack¹³ and OBF-1/BOB.1 (refs 14, 15). As RLIM contains a RING H2 zinc finger motif that is often found in ubiquitin protein ligases^{16,17}, we tested whether RLIM mediates E2-dependent ubiquitination of itself *in vitro*, as has been shown for other RING finger proteins¹⁸. Indeed, RLIM was ubiquitinated in the presence of the E2 ubiquitin-conjugating enzyme UbcH5 (Fig. 1a). This reaction was dependent on the RING finger because RING-finger-deleted or mutated versions of RLIM protein (RLIMΔRING and RLIMΔ9, respectively; see Methods) were no longer ubiquitinated by UbcH5 (Fig. 1b). As RLIM binds to nuclear LIM proteins and to CLIM cofactors⁸, we examined whether RLIM could serve as a ubiquitin protein ligase for these proteins. Whereas, under the conditions used, bacterially expressed full-length RLIM was not able, or only weakly able, to ubiquitinate the LIM-HD transcription factors Lhx1, Lhx3 and Isl1,

it efficiently mediated ubiquitination of LMO2 and LMO4 (Fig. 1c; see also Supplementary Information). The addition of either bacterially expressed full-length CLIM1 or CLIM2 inhibited the ubiquitination of LMO proteins. The presence of the LIM interaction domain in DN-CLIM, but not the dimerization-domain-containing CLIM Δ C, was required for this inhibition, suggesting a competition with RLIM for LIM domains (Fig. 1d). CLIM1 and CLIM2 were also poly-ubiquitinated in the presence of RLIM (Fig. 1e, f), and this reaction was not affected by adding bacterially expressed LMO2 or Lhx3 proteins (data not shown). Furthermore, RLIM was able to poly-ubiquitinate 35 S-labelled CLIM1 protein

complexed with Lhx3, which was purified in glutathione S-transferase (GST) pull-down experiments using bacterially expressed GST–Lhx3 protein (Fig. 1e). Bacterially expressed C-RLIM lacking the LIM and CLIM interaction domains was not able to promote either CLIM or LMO2 ubiquitinations (data not shown), indicating that these reactions are dependent on the presence of the LIM and CLIM interaction domains. Furthermore, the RING finger protein ligase Mdm2 did not ubiquitinate CLIM cofactors and LMO2 protein in similar experiments (data not shown), demonstrating the specificity of the RLIM-mediated ubiquitination reactions (see Supplementary Information). These results demonstrate that RLIM is a ubiquitin

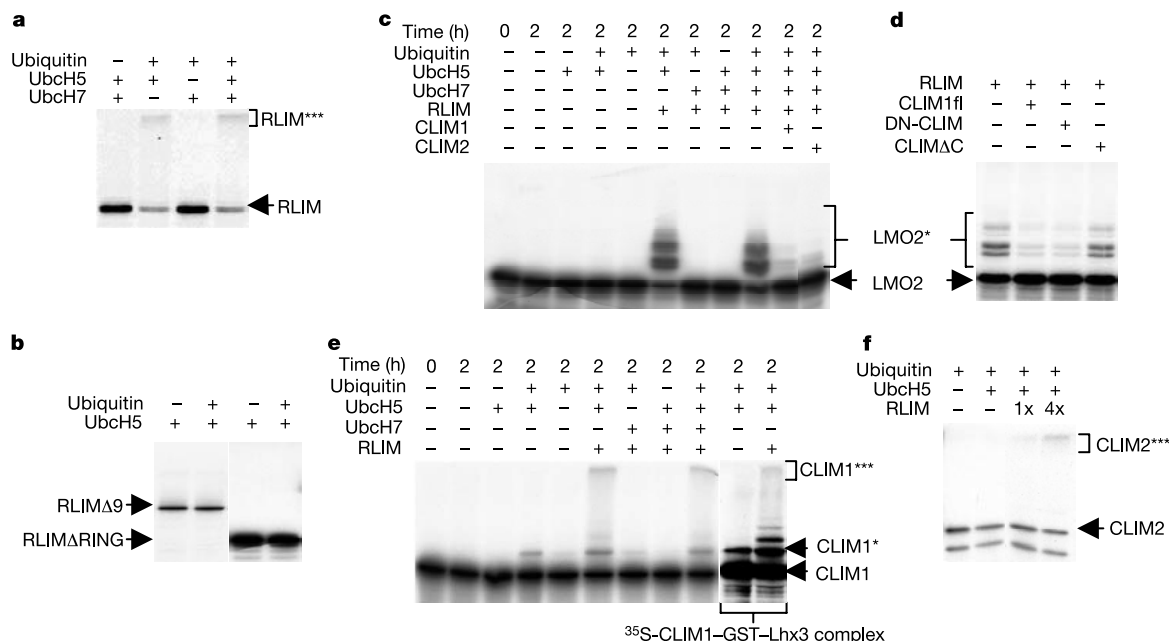


Figure 1 RLIM is a ubiquitin protein ligase that is able to ubiquitinate LMO proteins and CLIM cofactors *in vitro*. Poly-ubiquitinated proteins are indicated by three asterisks. **a**, 35 S-labelled RLIM protein incubated with ubiquitin-conjugating (E2) proteins Ubch5 and Ubch7. **b**, Ubiquitin protein ligase activity is abolished in RING finger deletion mutants RLIM Δ RING (residues 544–600 deleted) and RLIM Δ 9 (residues 562–571 deleted). **c**, Bacterially expressed RLIM ubiquitinates 35 S-labelled LMO2 proteins. Ubiquitinated

proteins are indicated by an asterisk. **d**, RLIM-mediated ubiquitination of 35 S-labelled LMO2 is inhibited by bacterially expressed full-length (fl) CLIM and DN-CLIM, but not CLIM Δ C. **e**, RLIM ubiquitinates CLIM1 cofactors. Putative mono-ubiquitinated CLIM1 is indicated by an asterisk. 35 S-labelled CLIM1 bound to GST–Lhx3 was purified in pull-down assays before the *in vitro* ubiquitination experiment. **f**, RLIM ubiquitinates CLIM2 cofactors in a dose-dependent manner.

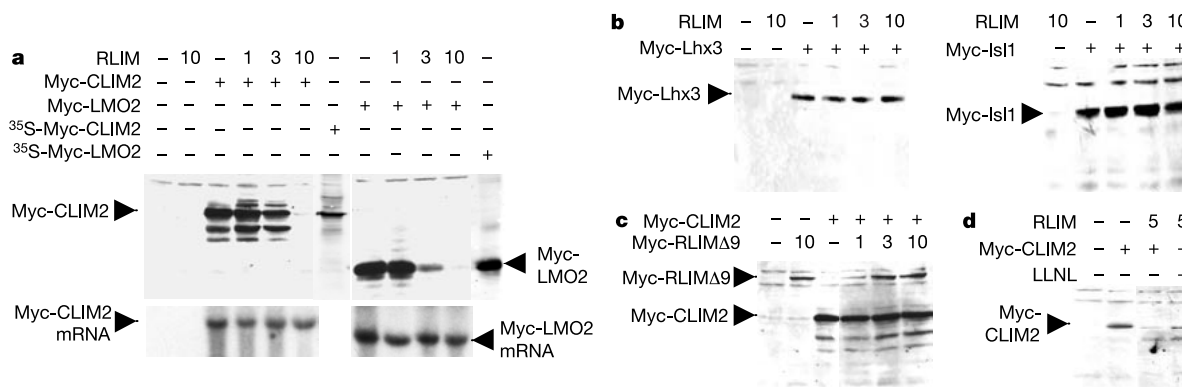


Figure 2 RLIM targets CLIM2 and LMO2 proteins for degradation. Western blots on total protein extracts of CHO cells co-transfected with the constructs indicated, using specific Myc antiserum, are shown. **a**, Extracts of cells co-transfected with non-tagged, full-length RLIM and Myc-CLIM2 or Myc-LMO2. *In vitro*-translated and 35 S-labelled Myc-tagged proteins are indicated. The messenger RNA levels in transfected cells were determined by northern blot using the 32 P-labelled 6x Myc complementary DNA probe. **b**, Extracts of

cells co-transfected with non-tagged full-length RLIM and Myc-Lhx3 or Myc-Is1 expression vectors. **c**, Extracts of cells co-transfected with Myc-tagged RLIM Δ 9 and Myc-CLIM2 expression vectors. **d**, Extracts of cells co-transfected with non-tagged, full-length RLIM and Myc-CLIM2 expression vector. Three hours before collection, cells were treated with 25 μ M of the proteasome inhibitor LLNL (MG101).

protein ligase that is able to specifically ubiquitinate LMO and CLIM proteins *in vitro*.

In transfected cells, overexpression of RLIM induced complete degradation of Myc-tagged CLIM2 and LMO2 in a dose-dependent manner (Fig. 2a), whereas the protein levels of LIM-HD transcription factors Isl1, Lhx1 and Lhx3 were only weakly affected (Fig. 2b; see also Supplementary Information). This degradation was dependent on a functional RING finger, because co-transfection of the Myc-RLIM Δ 9 construct abolished CLIM2 and LMO2 degradation (Fig. 2c and data not shown). Addition of the proteasome inhibitor LLNL to transfected cells partially impaired CLIM2 degradation, indicating proteolytic degradation through the 26S proteasome (Fig. 2d). To obtain evidence that endogenous RLIM is involved in endogenous CLIM degradation, we first compared cellular CLIM and RLIM protein levels. A reciprocal correlation between RLIM and CLIM protein levels was observed when comparing the gonadotrope pituitary cell line α T3 (ref. 19) to Chinese hamster ovary (CHO) cells and human embryonic kidney (HEK) 293 cells (Fig. 3a and data not shown), consistent with the idea that RLIM is involved in cellular CLIM degradation. Furthermore, when haemagglutinin (HA)-tagged full-length RLIM was transfected into the α T3 cell line, the endogenous CLIM levels (stained red) in the nuclei of transfected cells (stained green) dropped significantly (Fig. 3b). Conversely, in CHO, HEK 293 and ectodermal chick cells, endogenous CLIM levels increased on overexpression of the RLIM Δ 9

construct, presumably owing to its dominant negative effect¹⁷ on endogenous RLIM (Fig. 3c; see also Supplementary Information). These results were confirmed by RNA interference (RNAi) experiments in which the inhibition of endogenous RLIM expression caused an augmentation of CLIM levels (Fig. 3d). Together, these experiments demonstrate that endogenous RLIM is involved in CLIM degradation, thus placing RLIM in a central position for the developmental control of CLIM proteins.

To investigate the mechanisms by which RLIM regulates LIM-HD transcription-factors, we performed transient transfections. RLIM acts as a transcriptional co-repressor of Lhx3-mediated transcriptional activation from the α GSU and prolactin promoters (Fig. 4a, b) through the recruitment of Sin3A (ref. 8), a component of the histone deacetylase co-repressor complex. RLIM Δ 9 was able to inhibit CLIM-independent, but no longer CLIM-dependent, activation events mediated by Lhx3. Notably, RLIM Δ 9 interacted as efficiently as RLIM with Lhx3 and Sin3A (see Supplementary Information). These results indicate that RLIM can exert its repressive functions on LIM-HD transcription factors by two distinct but not mutually exclusive mechanisms, that is, by recruitment of Sin3A and by degradation of CLIM. We next examined the protein regions of RLIM responsible for protein-protein interactions with nuclear LIM domains and with CLIM cofactors⁸. Use of truncated RLIM proteins in a stringent GST protein interaction assay restricted this

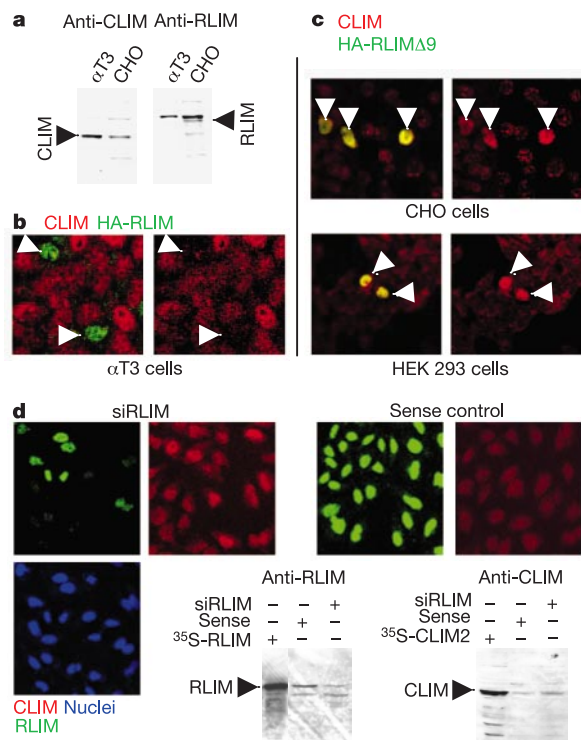


Figure 3 *In vivo* degradation of CLIM cofactors is mediated by RLIM. **a**, Western blots using CLIM and RLIM antisera showing reciprocal correlation of relative endogenous CLIM and RLIM protein levels in α T3 and CHO cell lines. **b**, Degradation of CLIM in transfected cells visualized by co-labelling with HA and CLIM antiserum. In α T3 cells, expression of HA-tagged, full-length RLIM (green) resulted in degradation of CLIM (red). **c**, Endogenous CLIM (red) levels augment on HA-RLIM Δ 9 (green, which combines with red to show yellow) transfections in CHO and HEK 293 cells. **d**, Inhibition of RLIM expression by RNAi entails increased CLIM levels. HeLa cells were transfected with double-stranded RNA oligonucleotides (siRLIM) or with a single-strand sense control. Cells were visualized with RLIM (green) or CLIM (red) antisera. An estimated 60–70% of the cells were transfected. In parallel experiments, protein levels were monitored by western blots. Independent experiments using a different siRLIM RNA led to similar results (not shown). Nuclei are visualized by blue staining.

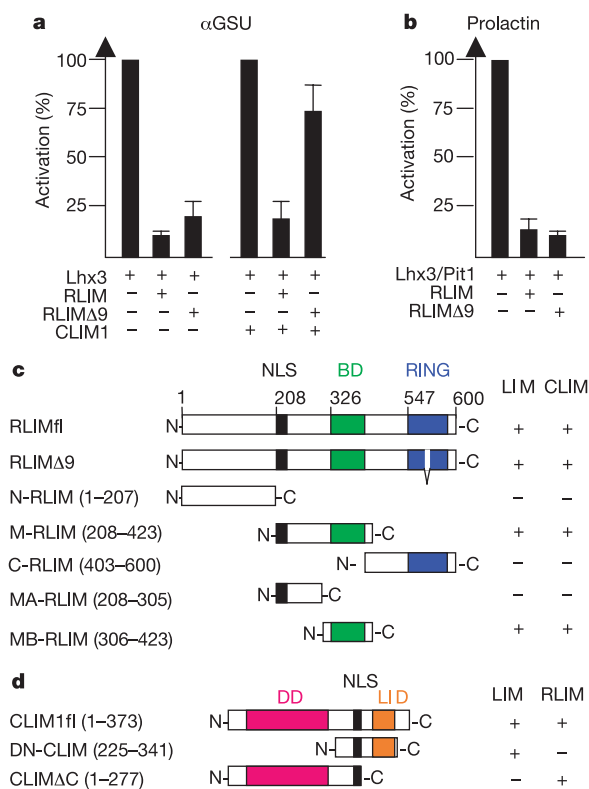


Figure 4 Functional protein interactions of RLIM. **a**, CHO cells were co-transfected with the α GSU promoter, and Lhx3 and LIM cofactor expression plasmids. CLIM co-transfections resulted in a 2.7-fold increase of activation levels compared with Lhx3 alone⁸. Lhx3/CLIM-mediated activation events are inhibited by RLIM in a RING-finger-dependent fashion. **b**, RLIM-mediated inhibition of the prolactin promoter is not dependent on the RING finger. CLIM co-transfections do not influence activities of this promoter⁸. **c**, Mapping of the LIM and CLIM interaction domains. ³⁵S-labelled *in vitro*-translated full-length RLIM and RLIM deletion mutants were tested for their ability to interact with GST fusion proteins. NLS, nuclear localization domain; BD, basic domain; RING, RING H2 zinc finger. **d**, Mapping of the RLIM interaction domain on CLIM. DD, dimerization domain; NLS, nuclear localization domain; LID, LIM-interaction domain.

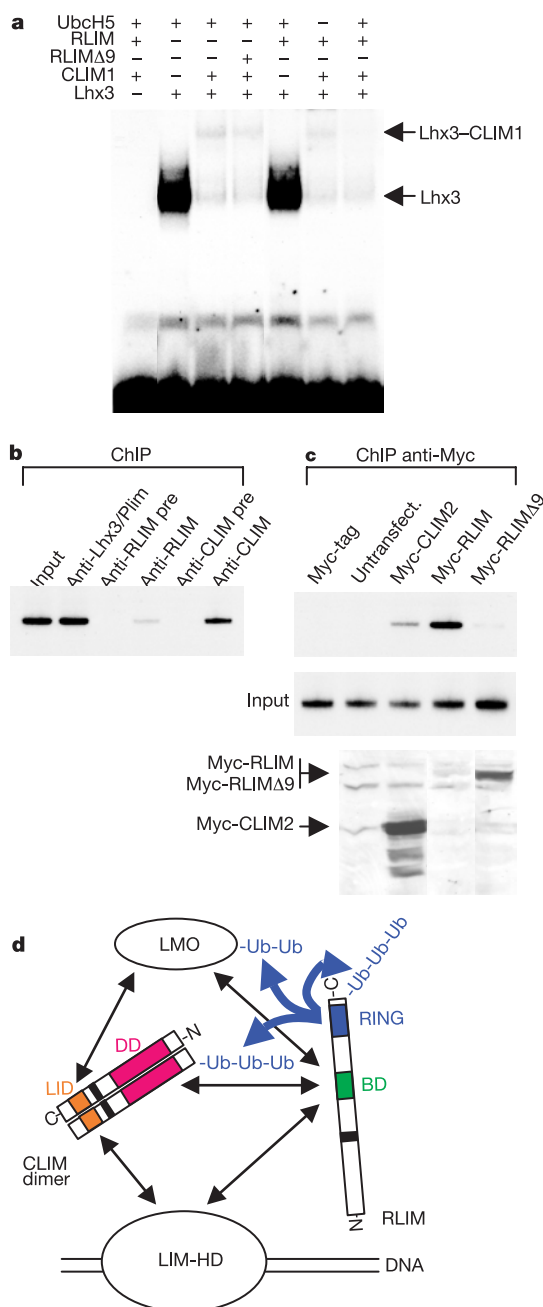


Figure 5 Cofactor exchange on LIM-HD proteins. **a**, Combined EMSA/*in vitro* ubiquitination assay demonstrating RLIM- and ubiquitination-dependent displacement of the DNA-bound Lhx3-CLIM1 complex. Poly-ubiquitinated CLIM-Lhx3-DNA complexes can probably not penetrate the gel. Experiments were performed using bacterially expressed cofactors CLIM1, RLIM and RLIMΔ9, and a ³²P-labelled, double-stranded oligonucleotide containing the Lhx3 binding sequence of the αGSU promoter in the presence of E1 and ubiquitin. **b**, Various quantities of Lhx3, CLIM, and RLIM are associated with the αGSU promoter *in vivo*. A ChIP assay on cell extracts of the αT3 cell line using anti-Lhx3/Plim, anti-RLIM, and anti-CLIM antisera is shown. pre, preimmune serum. **c**, ChIP assay on αT3 cells transfected with LIM cofactor expression constructs Myc-CLIM, Myc-RLIM and Myc-RLIMΔ9. The presence of the RING zinc finger greatly enhances the ability of RLIM to associate with this promoter. Protein levels in the same cell extracts were monitored in western blots using Myc-antiserum (lower panel). Lower Myc-RLIM levels probably reflect its auto-ubiquitination activity, resulting in a shorter half-life. **d**, Model of the regulation of LIM-HD transcription factors by cofactors CLIM and RLIM. Black arrows indicate protein-protein interactions; blue arrows indicate ubiquitination activity. Ubiquitinated proteins are shown with blue ubiquitin chains (Ub). By targeting LIM-domain-associated CLIM for degradation, DNA-bound LIM-HD transcription factors can subsequently interact with other protein partners, for example, RLIM.

interaction domain to a basic region of roughly 100 amino acids, conserved from chick to human²⁰. CLIM1ΔC, which harbours the dimerization domain, was sufficient to support RLIM interaction. DN-CLIM, containing the LIM interaction domain, did not interact (Fig. 4c, d; see also Supplementary Information).

CLIM cofactors are able to form complexes with LIM-HD proteins in electrophoretic mobility shift assay (EMSA) experiments, whereas this method does not support full-length RLIM binding to Lhx3 (refs 8, 21, 22). To investigate the effects of full-length RLIM on the DNA-bound Lhx3-CLIM complex, we performed a combined EMSA/*in vitro* ubiquitination assay in the presence of E1, UbcH5 and ubiquitin. Adding full-length RLIM to binding reactions resulted in a specific disappearance of the CLIM-Lhx3 complex (Fig. 5a). The fact that this effect required a functional RING finger and was dependent on UbcH5, indicates that RLIM is able to ubiquitinate CLIM cofactors associated with DNA-bound Lhx3 (see Supplementary Information). To obtain definitive evidence for the hypothesis that RLIM can associate with DNA-bound LIM-HD proteins by targeting CLIM for degradation, we tested the CLIM cofactor occupancy on the αGSU promoter using chromatin immunoprecipitation (ChIP). In agreement with the fact that αT3 cells synthesize αGSU¹⁹, the LIM-HD transcription factor Lhx3 and CLIM cofactors appeared to be predominantly associated with this promoter, whereas the association of RLIM seemed significantly lower (Fig. 5b). Examination of the cofactor occupancy of this promoter in transfected αT3 cells showed that Myc-CLIM2 protein associated with the promoter (Fig. 5c). We detected enhanced association of RLIM with the αGSU promoter in cells expressing full-length Myc-RLIM. This association was dependent on a functional RING finger, as the RLIMΔ9 protein—which binds equally well to Lhx3 when compared with wild-type protein (see Supplementary Information)—was not able to bind efficiently to the endogenous αGSU promoter.

The finding that RLIM is able to ubiquitinate CLIM cofactors bound to LIM-HD proteins, that CLIM cofactors can compete with RLIM for binding to LIM domains, and that a ubiquitination/degradation-dependent exchange in cofactor occupation on DNA occurs, provides an attractive mechanism for how cofactor exchange is mediated on LIM-HD proteins. Whether or not and in which order cofactor exchange occurs probably depends on LIM-HD, LMO, CLIM and RLIM nuclear concentrations, the ubiquitin-ligase activity of RLIM, and the availability of other LIM-domain interaction partners. This is demonstrated through the use of a model (Fig. 5d) showing the protein-protein interactions and ubiquitination events mediated by RLIM. As a growing number of transcription factor families have been shown to interact with several multiprotein complexes that confer multiple functions and activities^{1,2}, the degradation of selective protein members opens up intriguing possibilities of transcriptional regulation. Thus, regulated protein degradation together with competition for protein interaction domains probably represent general mechanisms used in combinatorial regulation of gene transcription. □

Methods

Antibodies and *in vitro* ubiquitination assays

Polyclonal antisera against mouse RLIM and CLIM were generated in rabbits (see Supplementary Information) and used for cofactor detection in western blots, immunocytochemistry and ChIP assays. Myc polyclonal antiserum and fluorescein isothiocyanate-labelled Myc and HA antisera (Santa Cruz) were used to detect Myc- and HA-tagged proteins, respectively. For RLIM-dependent ubiquitination, 2 μl of TNT rabbit reticulocyte lysate-translated proteins labelled with ³⁵S (Promega) were incubated in the presence of 50 ng E1, 50 ng each of bacterially expressed UbcH5 and/or UbcH7 and 6 μg ubiquitin (Sigma), and various bacterially expressed proteins, as indicated, in a reaction volume of 50 μl. In addition, the reactions contained 25 mM Tris-HCl at pH 7.5, 50 mM NaCl, 2 mM dithiothreitol, 4 mM ATP and 6 mM MgCl₂. After incubation for 2 h at 30 °C, the reactions were stopped, separated by SDS-polyacrylamide gel electrophoresis, and the ³⁵S-labelled proteins detected by fluorography.

Protein mutants and transient transfections

We used the following protein mutants: RLIM Δ 9 (residues 562–571 in the RING finger are deleted), RLIM Δ 2 Δ C (1–543), CLIM2 Δ C (1–271), Lhx3 Δ C (1–266), and proteins CLIM1, CLIM2 (ref. 21) and LMO2-LIM, RLIM, N-RLIM (1–207), M-RLIM (208–423), C-RLIM (403–600), DN-CLIM (225–341)⁸. Transient transfections and co-transfections were carried out using a Superfect transfection kit (Qiagen), as described previously²³.

Immunocytochemistry and RNAi

RNAi experiments were performed as reported²⁴, transfecting HeLa cells with double-stranded RNA oligonucleotides (Dharmacon) containing human RLIM sequences²⁰, or with a single-strand sense control (sequence information available on request) using the Oligofectamine kit (Invitrogen). Cells were visualized with RLIM or CLIM antisera and Toto-3 dye (Molecular Probes). In parallel experiments, protein levels of transfected cells were monitored by western blots using RLIM or CLIM antisera. Immunostainings of transfected cells were performed as reported²⁵. The experiments were analysed and pictures taken on a confocal microscope (Leica DMIRBE).

Protein–protein interactions and EMSA experiments

The *in vitro* protein–protein interaction experiments were carried out with bacterially expressed GST fusion proteins and ³⁵S-labelled proteins produced by *in vitro* transcription/translation using a TNT *in vitro* translation kit (Promega), as described²³, with the modification that all binding and washing steps were carried out at 37 °C. For ubiquitinations in EMSA, bacterially expressed proteins and ³²P-labelled oligonucleotides containing the LIM-HD recognition site in the α GSU promoter^{23,26} were incubated for 10 min on ice, as described^{8,21}, in buffer²⁷ supplemented with 4 mM ATP and 8 mM MgCl₂. Subsequently, E1, UbcH5 and ubiquitin were added, followed by a 2 h incubation at 30 °C.

Chromatin immunoprecipitation assays

ChIP assays were essentially done as described³, using anti-Lhx3/Plim²¹, anti-CLIM, anti-RLIM, and anti-Myc (BabCo) antisera, and oligonucleotides encompassing the region on the mouse α GSU promoter that contains the LIM-HD binding site^{23,26} (see Supplementary Information).

Received 12 October 2001; accepted 17 January 2002.

- Mannervik, M., Nibu, Y., Zhang, H. & Levine, M. Transcriptional coregulators in development. *Science* **284**, 606–609 (1999).
- Glass, C. K. & Rosenfeld, M. G. The coregulator exchange in transcriptional functions of nuclear receptors. *Genes Dev.* **14**, 121–141 (2000).
- Shang, Y., Hu, X., DiRenzo, J., Lazar, M. A. & Brown, M. Cofactor dynamics and sufficiency in estrogen receptor-regulated transcription. *Cell* **103**, 843–852 (2000).
- Bach, I. The LIM domain: regulation by association. *Mech. Dev.* **91**, 5–17 (2000).
- Hobert, O. & Westphal, H. Functions of LIM homeobox genes. *Trends Genet.* **16**, 75–83 (2000).
- Agulnick, A. D. *et al.* Interactions of the LIM-domain-binding factor Ldb1 with LIM homeodomain proteins. *Nature* **384**, 270–272 (1996).
- Morciello, P., Rosen, C., Baylies, M. K. & Dorsett, D. Chip, a widely expressed chromosomal protein required for segmentation and activity of a remote wing margin enhancer in *Drosophila*. *Genes Dev.* **11**, 2729–2740 (1997).
- Bach, I. *et al.* RLIM inhibits functional activity of LIM homeodomain transcription factors via recruitment of the histone deacetylase complex. *Nature Genet.* **22**, 394–399 (1999).
- Milán, M. & Cohen, S. M. Regulation of LIM homeodomain activity *in vivo*: a tetramer of dLDB and Apterous confers activity and capacity for regulation by dLMO. *Mol. Cell* **4**, 267–273 (1999).
- van Meyel, D. J. *et al.* Chip and Apterous physically interact to form a functional complex during *Drosophila* development. *Mol. Cell* **4**, 259–265 (1999).
- van Meyel, D. J. *et al.* Chip is an essential cofactor for Apterous in the regulation of axon guidance in *Drosophila*. *Development* **127**, 1823–1831 (2000).
- Milán, M., Diaz-Benjumea, F. J. & Cohen, S. M. Beadex encodes an LMO protein that regulates Apterous LIM-homeodomain activity in *Drosophila* wing development: a model for LMO oncogene function. *Genes Dev.* **12**, 2912–2920 (1998).
- Hershko, A. & Ciechanover, A. The ubiquitin system. *Annu. Rev. Biochem.* **67**, 425–479 (1998).
- Tiedt, R., Bartholdy, B. A., Matthias, G., Newell, J. W. & Matthias, P. The RING finger protein Siah-1 regulates the level of the transcriptional coactivator OBF-1. *EMBO J.* **20**, 4143–4152 (2001).
- Boehm, J., He, Y., Greiner, A., Staudt, L. & Wirth, T. Regulation of BOB-1/OBF-1 stability by SIAH. *EMBO J.* **20**, 4153–4162 (2001).
- Joazeiro, C. A. P. & Weissman, A. M. RING finger proteins: mediators of ubiquitin ligase activity. *Cell* **79**, 549–552 (2000).
- Pickart, C. M. Mechanisms underlying ubiquitination. *Annu. Rev. Biochem.* **70**, 503–533 (2001).
- Loricke, K. L. *et al.* RING fingers mediate ubiquitin-conjugating enzyme (E2)-dependent ubiquitination. *Proc. Natl Acad. Sci. USA* **96**, 11364–11369 (1999).
- Windle, J. J., Weiner, R. I. & Mellon, P. L. Cell lines of the pituitary gonadotrope lineage derived by targeted oncogenesis in transgenic mice. *Mol. Endocrinol.* **4**, 597–603 (1990).
- Ostendorff, H. P. *et al.* Functional characterization of the gene encoding RLIM, the corepressor of LIM homeodomain factors. *Genomics* **69**, 120–130 (2000).
- Bach, I., Carrière, C., Ostendorff, H. P., Andersen, B. & Rosenfeld, M. G. A family of LIM domain interacting cofactors confer transcriptional synergism between LIM and Otx homeoproteins. *Genes Dev.* **11**, 1370–1380 (1997).
- Jurata, L. W. & Gill, G. N. Functional analysis of the nuclear LIM domain interactor NLI. *Mol. Cell Biol.* **17**, 5688–5698 (1997).
- Bach, I. *et al.* P-Lim, a LIM homeodomain factor, is expressed during pituitary organ and cell commitment and synergizes with Pit-1. *Proc. Natl Acad. Sci. USA* **92**, 2720–2724 (1995).
- Elbashir, S. M. *et al.* Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature* **411**, 494–498 (2001).

- Bach, I. & Yaniv, M. More potent transcriptional activators or a transdominant inhibitor of the HNF1 homeoprotein family are generated by alternative RNA processing. *EMBO J.* **12**, 4229–4242 (1993).
- Roberson, M. S., Schoderbek, W. E., Tremml, G. & Maurer, R. A. Activation of the glycoprotein hormone α -subunit promoter by a LIM-homeodomain transcription factor. *Mol. Cell Biol.* **14**, 2985–2993 (1994).
- Yu, V. C. *et al.* RXR β : a coregulator that enhances binding of retinoic acid, thyroid hormone, and vitamin D receptors to their cognate response elements. *Cell* **67**, 1251–1266 (1991).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com>).

Acknowledgements

We thank P. Mellon for the gift of α T3 cells; A. Krones and Y. Alvarez for advice on ChIP assays and chick electroporations, respectively; T. Heinzel for Sin3 expression vectors; I. Hermans-Borgmeyer for comments on the manuscript; and Y. Bodingbauer for technical assistance. H.P.O. was a fellow of the Graduiertenkolleg 255 of the University of Hamburg, and I.B. is a Heisenberg scholar of the Deutsche Forschungsgemeinschaft (DFG). I.B. and M.S. are supported by grants from the DFG.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to I.B. (e-mail: ingolf.bach@zmnh.uni-hamburg.de).

Structure of the HP1 chromodomain bound to histone H3 methylated at lysine 9

Peter R. Nielsen*, Daniel Nietlispach*, Helen R. Mott*, Juliana Callaghan*, Andrew Bannister†, Tony Kouzarides†, Alexey G. Murzin‡, Natalia V. Murzina* & Ernest D. Laue*

* Cambridge Centre for Molecular Recognition, Department of Biochemistry, University of Cambridge, 80 Tennis Court Road, Cambridge CB2 1GA, UK
† Wellcome/Cancer Research UK, Institute of Cancer and Developmental Biology and Department of Pathology, University of Cambridge, Tennis Court Road, Cambridge CB2 1QR, UK
‡ MRC Centre for Protein Engineering, Hills Road, Cambridge CB2 2QH, UK

Specific modifications to histones are essential epigenetic markers¹—heritable changes in gene expression that do not affect the DNA sequence. Methylation of lysine 9 in histone H3 is recognized by heterochromatin protein 1 (HP1), which directs the binding of other proteins to control chromatin structure and gene expression^{2–4}. Here we show that HP1 uses an induced-fit mechanism for recognition of this modification, as revealed by the structure of its chromodomain bound to a histone H3 peptide dimethylated at N ϵ of lysine 9. The binding pocket for the N-methyl groups is provided by three aromatic side chains, Tyr 21, Trp 42 and Phe 45, which reside in two regions that become ordered on binding of the peptide. The side chain of Lys 9 is almost fully extended and surrounded by residues that are conserved in many other chromodomains. The QTAR peptide sequence preceding Lys 9 makes most of the additional interactions with the chromodomain, with HP1 residues Val 23, Leu 40, Trp 42, Leu 58 and Cys 60 appearing to be a major determinant of specificity by binding the key buried Ala 7. These findings predict which other chromodomains will bind methylated proteins and suggest a motif that they recognize.

Histone methylation is important in the regulation of chromatin structure and thus gene expression⁵. Methylation at lysine 9 of histone H3 by Suv39 methyltransferase⁶ and its relatives strongly

correlates with a repressed chromatin state^{7,8}. Transcriptional repression through Lys 9 methylation has been demonstrated at centromeric heterochromatin in yeast^{2,4} and at retinoblastoma-regulated promoters in mammalian cells⁹. This repression requires the specific recognition of Lys 9-methylated histone H3 by the HP1 chromodomain^{2,3}.

The structure of the free chromodomain of mouse HP1 β (also known as MOD1 or M31)¹⁰ provided no clue as to the mechanism of its recognition of methyl lysine. Methylation of DNA is known to control protein–nucleic acid interactions in many biological systems, but how protein methylation promotes protein–protein interactions is unknown. There is no precedent for the molecular recognition of a lysine residue modified with a variable number of extra methyl groups at the end of its flexible side chain. Unlike acetylation, methylation neither introduces new hydrogen-bonding opportunities nor changes the net charge, so there can be no contribution from additional hydrogen bonds and/or electrostatic interactions. The mechanism of methyl lysine recognition by the HP1 chromodomain thus has wide implications for our understanding of molecular and structural biology. To understand methyl lysine recognition, and to illuminate the role of other chromodomains, we determined the structure of the HP1 β chromodomain–histone H3 peptide complex (Fig. 1).

In the complex, the chromodomain adopts the canonical fold found previously in the free and shadow chromodomains^{10–13}. The histone H3 peptide binds in an extended β -strand-like conformation in the groove predicted previously¹⁰, using an induced-fit mechanism for recognition of the methylated peptide (Fig. 1a, b). The amino-terminal tail of the free chromodomain wraps around the peptide as it binds (Fig. 1c). This brings together the side chains of three conserved aromatic residues, Tyr 21, Trp 42 and Phe 45, which bind the methyl groups of dimethyl lysine 9 in a pocket formed by their mutually orthogonal aromatic rings (Fig. 2a). The positions of the methyl lysine and interacting chromodomain

Table 1 Experimental restraints and structural statistics

Number of experimental restraints	Unambiguous	Ambiguous
Distance restraints from NOEs		
Intramolecular (HP1 β)	859	1,740
Intramolecular (peptide)	25	17
Intermolecular	30	188
Dihedral restraints ²⁴ (HP1 β)	22	
	$\langle SA \rangle$	$\langle SA \rangle_c$
Coordinate precision*		
RMSD of backbone atoms (Å)	0.68 $\pm$ 0.07	0.54
RMSD of all heavy atoms (Å)	1.15 $\pm$ 0.08	1.07
RMS deviations		
From the experimental restraints		
NOE distances (Å)	0.018 $\pm$ 0.0047	0.024
Dihedral angles (°)	0.386 $\pm$ 0.106	0.331
From idealized geometry		
Bonds (Å)	0.002 $\pm$ 0.00014	0.0017
Angles (°)	0.366 $\pm$ 0.0082	0.372
Impropers (°)	0.248 $\pm$ 0.0154	0.260
Final energy, E_{LJ} (kJ mol ⁻¹)†	-282.5 $\pm$ 17	-266.41
Ramachandran analysis		
Residues in most-favoured regions	74.1%	80.4%
Residues in additionally allowed regions	19.7%	17.9%
Residues in generously allowed regions	4.1%	0.0%
Residues in disallowed regions	2.1%	1.8%

$\langle SA \rangle$ is the average root mean square (RMS) deviation for the ensemble; $\langle SA \rangle_c$ is the value for the structure that is closest to the mean. RMSD, RMS deviation. Confidence intervals are s.d.

*Computed over residues 19–71 of HP1 β and 4–10 of the histone H3 peptide.

†The Lennard-Jones potential was not used at any stage in the refinement.

residues are all well determined by the nuclear magnetic resonance (NMR) data, but there is space for one more methyl group in this pocket, and trimethyl lysine binds similarly (the dissociation constant K_d for the dimethylated and trimethylated peptides are 2.1 and 1.9 μ M, respectively). It is significant that aromatic rather than aliphatic residues are used for the recognition of the methyl groups. Without the methyl groups, the rigid side chains of the aromatic

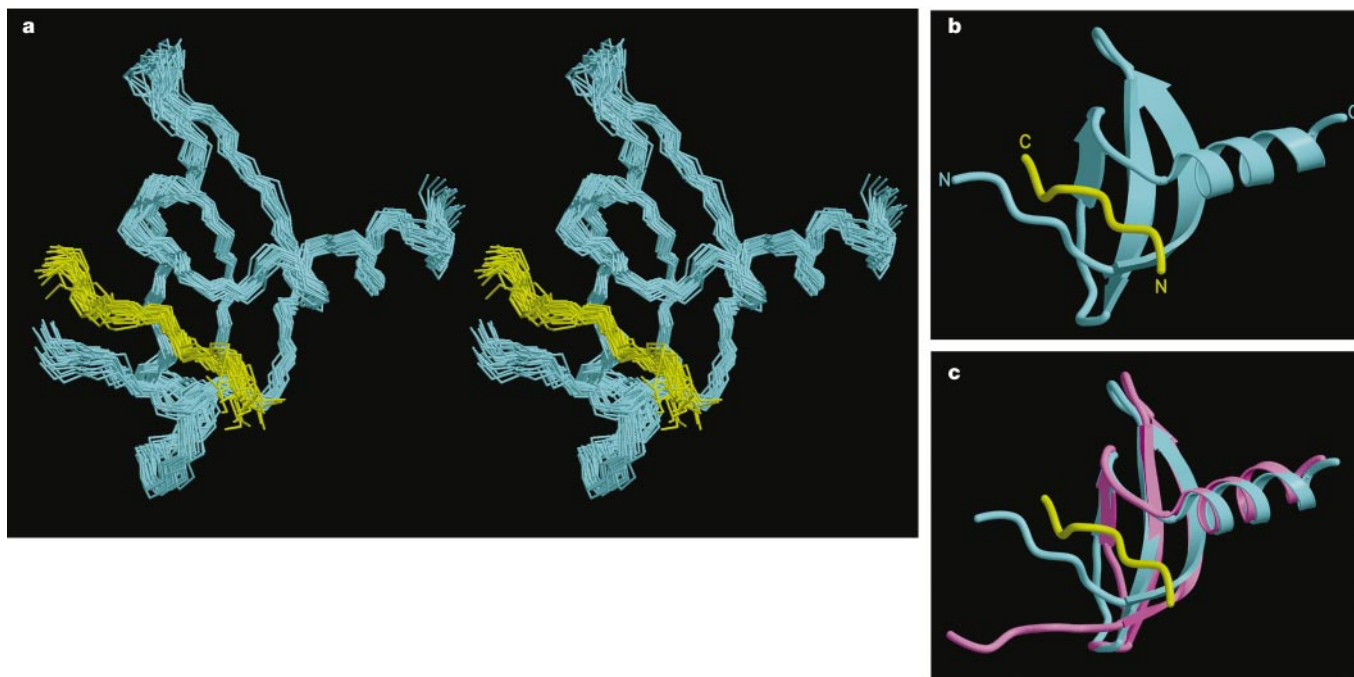


Figure 1 Structure of the HP1 β chromodomain complexed with a histone H3 peptide dimethylated at lysines 4 and 9. **a**, Stereo view of the backbone of residues 19–71 of HP1 β and 4–10 of histone H3 from the 25 lowest energy structures (of 75 that converged from the 100 computed). In the final structures no distance restraints were violated by more than 0.5 Å and no dihedral angle restraints by more than 5°. The structure has good

covalent geometry and non-bonded contacts (Table 1). **b**, The structure closest to the mean in the same orientation as that shown in **a**. HP1 β is blue and the histone H3 peptide is yellow. **c**, The structure closest to the mean overlaid with that of the free HP1 β chromodomain¹⁰ in pink. Figures 1 and 2 were generated using Molscript²⁵ and Raster 3D²⁶.

residues probably can not form the tight cluster that stabilizes the peptide-binding conformation. In contrast, more flexible aliphatic side chains would pack together without forming a pocket. Moreover, with aromatic residues interactions between π electrons and the positively charged nitrogen can contribute to the binding. This hypothesis is supported by the observation that the Y21L mutation significantly reduces peptide binding (data not shown). These aromatic residues are conserved in other chromodomains, as are Thr 51, which interacts with the Lys 9 side chain, and Glu 53 and Asn 57, which interact with the peptide backbone at the other end of the methyl lysine binding site (Fig. 2a). These two sets of residues thus define the methyl lysine binding site, which binds in an almost fully extended conformation along Trp 42 and Thr 51. Specificity is provided because no other hydrophobic side chain would be large enough to span the distance between the methyl-binding pocket and the backbone-binding site, and thus be able to substitute for methyl lysine (Fig. 2a).

The structure and proposed mechanism of methyl lysine recognition allows us to predict which chromodomain-like sequences adopt the same fold and probably bind peptides that contain methyl lysine (Fig. 3a). The residue at position 42 (invariably tryptophan) is in the hydrophobic core and is probably essential for the chromodomain fold. However, there are no specific structural requirements for aromatic residues at positions 21 and 45, although both are

found in most chromodomains, including those of the HP1, the Suv39 and Pc-like families, chromomethylase, *Tetrahymena thermophila* Pdd1p, and other proteins (Fig. 3a). The other residues that interact with methyl lysine—Thr 51, Glu 53 and Asn 57 in HP1—are also generally conserved in this set of chromodomains, which supports our predictions (Fig. 3a). However, there are occasional, non-essential variations of these residues in some proteins, for example, N57D in the HP1 fission yeast homologue SWI6, which is known to bind Lys 9-methylated H3 (refs 2, 4). The sequences of the CHD proteins suggest that only the first of the pair of chromodomains in CHD1 and CHD2 could recognize methyl lysine, whereas neither of the chromodomains in CHD3 and CHD4 conform with the requirements for methyl lysine recognition (Fig. 3a).

In other chromodomains, which lack the complete set of conserved aromatic residues, there is no conservation of the other methyl lysine-interacting residues. Most of these sequences are compatible with the structural determinants of the chromodomain fold, including several proteins that have no known chromatin connection (Fig. 3a). The only exceptions are the 'RNA-binding chromodomains' from MSL3, MOF1 and related proteins¹⁴. These do not contain the methyl lysine-interacting residues, but they are also unlikely to adopt the canonical chromodomain fold because they lack principal conserved hydrophobic amino acids (Fig. 3a).

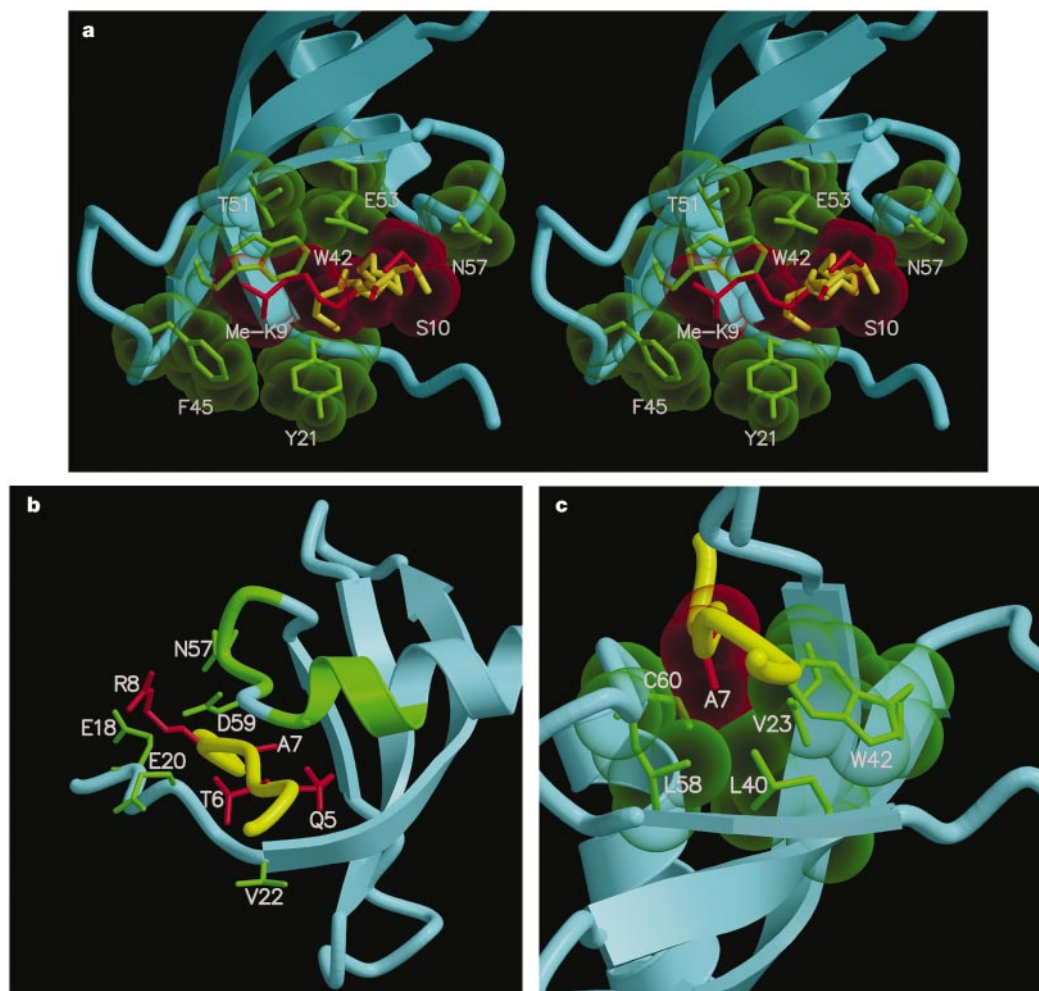


Figure 2 The interface between the chromodomain and the methylated peptide. **a**, Tyr 21, Trp 42 and Phe 45 form a 'hydrophobic box' that binds the *N*-methyl groups. The methyl lysine side chain binds along the surface of Trp 42 and Thr 51, whereas Glu 53 and Asn 57 contact the C α end. **b**, Gln 5 interacts with the N terminus of the α -helix and Arg 8

interacts with a patch of acidic residues to help structure the flexible N-terminal tail. **c**, Ala 7 binds in the core of the chromodomain, making contacts with residues Val 23, Leu 40, Trp 42, Leu 58 and Cys 60. In **a**, **b** and **c**, side chains of the histone H3 peptide are red and interacting residues in the chromodomain are green.

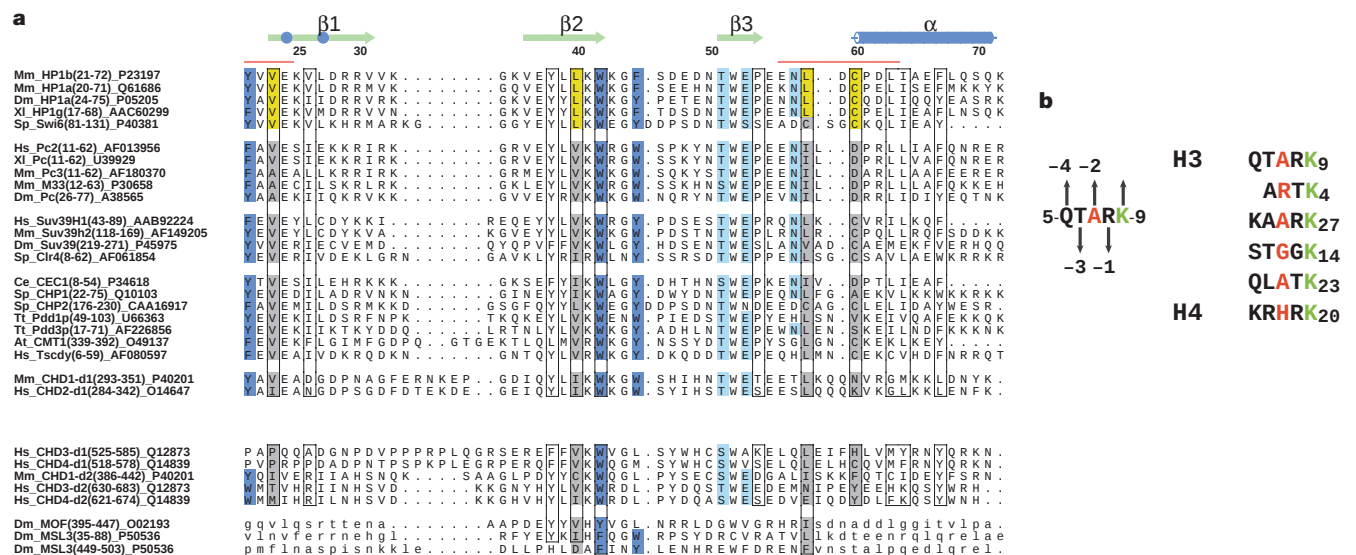


Figure 3 Sequence comparisons and residues involved in complex formation. **a**, Chromodomain alignment numbered for mouse HP1 β . Residues important for methyl lysine recognition are blue (N-methyl groups) and light blue (Lys 9 side chain or backbone). Residues that make contacts with the K - 2 position of the peptide (see text) are shown in yellow (HP1 family) or grey (others). The two segments of HP1 β that form the walls of the peptide-binding groove are indicated by red lines. β -strands are indicated by green arrows and the α -helix by a blue cylinder. The blue dots indicate the positions of conserved bulges in β 1. Conserved hydrophobic residues that define the fold and the borders of the turn between the β -strand and the α -helix (Pro 54 and Cys 60, respectively) are boxed. The regions in MSL3 and MOF1 that do not conform with a

canonical chromodomain fold are shown in small letters. Sequences are labelled by species name (Mm, *Mus musculus*; Dm, *Drosophila melanogaster*; Xl, *Xenopus laevis*; Hs, *Homo sapiens*; Sp, *Schizosaccharomyces pombe*; Ce, *Caenorhabditis elegans*; Tt, *Tetrahymena thermophila*; At, *Arabidopsis thaliana*), followed by protein name, residue range and GenBank accession number. Produced using Alscript²⁷. **b**, The sequence of the N-terminal tail of histone H3. Up arrows denote side chains that point into the chromodomain core; down arrows denote surface side chains. The right panel shows the N-terminal histone H3 and H4 sequences aligned around the lysines. The key K - 2 position is shown in red and the lysines in green.

The HP1 chromodomain does not recognize methyl lysine as a free amino acid¹⁵. We also found that neither of its derivatives, 4-dimethylamino-butyric acid (pH 7.5) and 3-dimethylamino-1-propanol, altered the NMR spectra when added to the HP1 β chromodomain (data not shown). The recognition of methyl lysine thus depends on the context in which it is presented. The peptide we used was dimethylated at both Lys 4 and Lys 9, but only the dimethyl group of Lys 9 is buried in the structure. Consistent with this, a Lys 9-methylated peptide binds as well (K_d 0.7 μ M) as the Lys 4,9 bis-methylated peptide (K_d 1.9 μ M). Moreover, the complexes have virtually identical NMR spectra. The structure thus confirms that methylation of Lys 4 neither contributes to nor interferes with recognition of methylated Lys 9.

The additional contacts to the chromodomain mainly involve the four amino acids that precede Lys 9: QTAR (Figs 2b and 3b). The structure allows us to propose a peptide motif that is recognized by the HP1 chromodomains. The side chain in the -2 position relative to methyl lysine (K - 2) seems to be central for specificity: Ala 7 is buried in a small hydrophobic pocket, formed by residues Val 23, Leu 40, Trp 42, Leu 58 and Cys 60, which are conserved in the HP1 family (Fig. 3a). Only alanine, or the smaller glycine, can be accommodated (Fig. 2c). Disruption of this binding site by mutation of Val 23 in *Drosophila* HP1 to the bulkier methionine abolishes binding of Lys 9-methylated H3 (refs 2, 3) and impairs its biological function¹⁶. At K - 4, Gln 5 interacts with the N terminus of the chromodomain α -helix (Fig. 2b). Here, positively charged residues would prevent binding owing to electrostatic repulsion by the helix dipole, whereas bulky hydrophobic residues will obstruct binding owing to interference with the hydration of the helix N terminus. The K - 1 and K - 3 side chains (Thr 6 and Arg 8) point away from the interaction surface. It is probable that any amino acid could be tolerated at these positions, but Arg 8 interacts with acidic residues in the flexible N-terminal tail (Fig. 2b) and probably contributes to binding.

Other lysines are methylated in histone tails, namely Lys 4 and

Lys 27 in H3, and Lys 20 in H4 (Fig. 3b). Lys 4-methylated H3, which does not bind HP1, has an unacceptable arginine at K - 2, as does Lys 20-methylated H4 (histidine). Lys 27-methylated H3 has an acceptable alanine at K - 2, but an unfavourable positively charged and bulky lysine at K - 4. This explains why Lys 27-methylated H3 binds HP1 much more weakly, despite its similar sequence (data not shown). Of the other lysines, only Lys 14 and Lys 23 in H3 would present the modification in a suitable context for recognition by the HP1 chromodomain. However, neither is known to be methylated *in vivo*, although both can be acetylated.

Residues in HP1 β that interact with Lys 9-methylated histone H3 are conserved in the HP1 family, for example, those interacting with Ala 7 (see above), but not in the other chromodomains. There is, however, conservation of the equivalent residues in separate chromodomain families and subfamilies, suggesting that each may recognize methyl lysine in a different, family-specific context (Fig. 3a). For example, the Pc-like chromodomains are likely to have a bigger pocket for the residue at K - 2. It is also possible that the binding of methyl lysine-containing peptides to other chromodomains might occur in the reverse orientation to that observed in HP1. If this were the case, the key positions for context recognition would be K + 2 and K + 4 rather than K - 2 and K - 4. As new methylation sites are identified, we will be able to refine our understanding of how protein methylation regulates chromodomain binding. □

Methods

Expression and purification of the HP1 chromo domain

Unlabelled, and ¹⁵N- and ¹³C-labelled mouse HP1 β chromodomains (residues 10-80) were expressed and purified as described¹¹.

Synthetic peptides and binding assays

The synthetic N-terminal histone H3 peptides (NH₂-ARTK(Me)₂QTARK(Me)₂STGGKAPGG-COOH; NH₂-ARTK(Me)₃QTARK(Me)₃STGGKAPGG-COOH and NH₂-ARTKQTARK(Me)₃STGGKAPGG-COOH) were synthesized and purified by G. Bloomberg. As judged by measurement of changes in the chromodomain's intrinsic tryptophan

fluorescence, these peptides bind with affinities of 2.14 ± 0.36 , 1.94 ± 0.65 and $0.68 \pm 0.05 \mu\text{M}$, respectively.

Expression, purification and methylation of the peptide

Glutathione S-transferase (GST)–H3 (residues 1–13) was expressed in *Escherichia coli* BL21 cells and purified using standard protocols. The GST-fusion peptide was digested with thrombin and purified by reverse phase HPLC on an octadecyl (C_{18}) column using a buffer containing 0.1% trifluoroacetic acid in acetonitrile/water. The eluted peptide was freeze dried, dissolved in water and dimethylated as described¹⁷ using a 20% ^{13}C -labelled formaldehyde (CIL)/water solution. The uniform methylation of the peptide was verified by mass spectrometry.

NMR spectroscopy

Purified chromodomain, after gel filtration chromatography, was concentrated to about 200 μl and mixed directly with histone H3 peptide to make NMR samples in 150 mM NaCl, 20 mM phosphate buffer (either pH 5.5 or 7.5), 10% D_2O . The peptide binds in fast exchange on the NMR timescale and spectra were recorded (at 25 °C on Bruker DRX600 and DRX800 spectrometers) using 2:1 and 0.9:1 (peptide:protein) ratios for assignment of the protein and peptide resonances, respectively. The protein backbone and side-chain resonances were assigned with standard triple-resonance NMR techniques¹⁸ using a $^{13}\text{C}/^{15}\text{N}$ -labelled sample of HP1 β chromodomain complexed with the unlabelled 18-residue histone H3 peptide. The peptide resonances were assigned using ^{13}C , ^{15}N -rejected nuclear Overhauser effect and total correlation spectroscopy (NOESY and TOCSY) experiments and the methyl group resonances of lysines 4 and 9 were confirmed with the help of a sample comprising an unlabelled chromodomain complexed to a selectively ^{13}C -methyl-labelled histone H3 peptide. NOEs were determined using two-dimensional NOESY and three-dimensional ^{13}C - and ^{15}N -separated NOESY spectra. Intermolecular contacts were obtained from a $^{13}\text{C}/^{15}\text{N}$ X-filtered NOESY spectrum¹⁹ and a three-dimensional J(CH,NH)-separated NOESY spectrum²⁰. Spectra were processed with the AZARA suite of programs (W. Boucher, unpublished) and analysed with ANSIG²¹.

Structure determination

Structures were calculated from an extended template using simulated annealing in CNS version 1.0 (ref. 22) and ARIA²³, which uses the previous structures to identify restraint violations and removes ambiguous restraints that do not contribute sufficiently to the intensity of the peak. To generate the final NOE tables, nine iterations were calculated, each on 20 structures. Finally, 100 structures were calculated, from which the 25 with lowest energy were selected.

Received 9 January; accepted 7 February 2002.

Published online 20 February 2002, DOI 10.1038/nature722.

1. Strahl, B. D. & Allis, C. D. The language of covalent histone modifications. *Nature* **403**, 41–45 (2000).
2. Bannister, A. J. *et al.* Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromodomain. *Nature* **410**, 120–124 (2001).
3. Lachner, M., O'Carroll, D., Rea, S., Mechtler, K. & Jenuwein, T. Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature* **410**, 116–120 (2001).
4. Nakayama, J., Rice, J. C., Strahl, B. D., Allis, C. D. & Grewal, S. I. Role of histone H3 lysine 9 methylation in epigenetic control of heterochromatin assembly. *Science* **292**, 110–113 (2001).
5. Zhang, Y. & Reinberg, D. Transcription regulation by histone methylation: interplay between different covalent modifications of the core histone tails. *Genes Dev.* **15**, 2343–2360 (2001).
6. Rea, S. *et al.* Regulation of chromatin structure by site-specific histone H3 methyltransferases. *Nature* **406**, 593–599 (2000).
7. Noma, K., Allis, C. D. & Grewal, S. I. Transitions in distinct histone H3 methylation patterns at the heterochromatin domain boundaries. *Science* **293**, 1150–1155 (2001).
8. Litt, M. D., Simpson, M., Gaszner, M., Allis, C. D. & Felsenfeld, G. Correlation between histone lysine methylation and developmental changes at the chicken β -globin locus. *Science* **293**, 2453–2455 (2001).
9. Nielsen, S. J. *et al.* Rb targets histone H3 methylation and HP1 to promoters. *Nature* **412**, 561–565 (2001).

10. Ball, L. J. *et al.* Structure of the chromatin binding (chromo) domain from mouse modifier protein 1. *EMBO J.* **16**, 2473–2481 (1997).
11. Brasher, S. V. *et al.* The structure of mouse HP1 suggests a unique mode of single peptide recognition by the shadow chromo domain dimer. *EMBO J.* **19**, 1587–1597 (2000).
12. Horita, D. A., Ivanova, A. V., Altieri, A. S., Klar, A. J. & Byrd, R. A. Solution structure, domain features, and structural implications of mutants of the chromo domain from the fission yeast histone methyltransferase Clr4. *J. Mol. Biol.* **307**, 861–870 (2001).
13. Cowieson, N. P., Partridge, J. F., Allshire, R. C. & McLaughlin, P. J. Dimerisation of a chromo shadow domain and distinctions from the chromodomain as revealed by structural analysis. *Curr. Biol.* **10**, 517–525 (2000).
14. Akhtar, A., Zink, D. & Becker, P. B. Chromodomains are protein–RNA interaction modules. *Nature* **407**, 405–409 (2000).
15. Jacobs, S. A. *et al.* Specificity of the HP1 chromo domain for the methylated N-terminus of histone H3. *EMBO J.* **20**, 5232–5241 (2001).
16. Platero, J. S., Hartnett, T. & Eisenberg, J. C. Functional analysis of the chromo domain of HP1. *EMBO J.* **14**, 3977–3986 (1995).
17. Means, G. E. & Feeney, R. E. Reductive alkylation of amino groups in proteins. *Biochemistry* **7**, 2192–2201 (1968).
18. Fierentz, A. E. & Wagner, G. NMR: A multifaceted approach to macromolecular structure. *Q. Rev. Biophys.* **33**, 29–65 (2000).
19. Zwahlen, C. *et al.* Methods for measurement of intermolecular NOEs by multinuclear NMR spectroscopy: Application to a bacteriophage λ N-peptide/boxB RNA complex. *J. Am. Chem. Soc.* **119**, 6711–6721 (1997).
20. Melacini, G. Separation of intra- and intermolecular NOEs through simultaneous editing and J-compensated filtering: A quadrature-free constant-time J-resolved approach. *J. Am. Chem. Soc.* **122**, 9735–9738 (2000).
21. Kraulis, P. J. *et al.* The solution structure and dynamics of ras p21.GDP determined by heteronuclear three- and four-dimensional NMR spectroscopy. *Biochemistry* **33**, 3515–3531 (1994).
22. Brunger, A. T. *et al.* Crystallography and NMR system (CNS): A new software system for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
23. Linge, J. P., O'Donoghue, S. J. & Nilges, M. Automated assignment of ambiguous NOEs with ARIA. *Methods Enzymol.* **339**, 71–90 (2001).
24. Cornilescu, G., Delaglio, F. & Bax, A. Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J. Biol. NMR* **13**, 289–302 (1999).
25. Kraulis, P. J. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24**, 946–950 (1991).
26. Merrit, E. A. & Murphy, M. E. P. Raster3D Version 2.0. A program for photorealistic molecular graphics. *Acta Crystallogr. D* **50**, 869–873 (1994).
27. Barton, G. J. ALSCRIPT: a tool to format multiple sequence alignments. *Protein Eng.* **6**, 37–40 (1993).

Acknowledgements

We thank A. Verreault for the pGEX-H3 construct; R. Turner, P. Sharratt and the PNAC facility for mass spectrometry and amino-acid analysis; A. Thiru for comments on the manuscript; the Medical Research Council for a Career Development Award (to H.R.M.); and the Wellcome Trust for financial support. The Cambridge Centre for Molecular Recognition is supported by the Biotechnology and Biological Sciences Research Council and the Wellcome Trust.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for material should be addressed to N.V.M. (e-mail: nm@mole.bio.cam.ac.uk) or E.D.L. (e-mail: e.d.laue@bioc.cam.ac.uk). The coordinates have been deposited in the Protein Data Bank under accession number 1GUW.

Contacts

Publisher: Fabien Savenay
Editor: Paul Smaglik
Sales Director: Ben Crowe

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Group European Manager:

Leonie Weiss (4954)

European Manager:

Nevin Bayoumi (4978)

UK/ RoW/ Ireland:

Matt Powell (4953), Ben Corp (4974),
Andy Douglas (4975)

Holland/ Italy:

Nevin Bayoumi (4978)

Scandinavia:

Sille Opstrup (4994)

Spain/ Portugal:

Leonie Weiss (4954)

Production Manager:

Billie Franklin

To send films and materials use

London address above.

Tel +44 (0) 20 7843 4814

Fax +44 (0) 20 7843 4996

e-mail: naturejobs@nature.com

International

Advertising Coordinator:

Laura Pearson (4977)

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Ben Lund

European Satellite Offices

France/ Belgium:

Christine Niox-Chateau

Tel + 33 (0) 1 43 20 16 51

Fax + 33 (0) 1 43 20 51 52

e-mail: c.nioxchateau@nature.com

Germany/ Austria/ Switzerland:

Patrick Phelan/ Kate Turner

Tel + 49 89 54 90 57 11/2

Fax + 49 89 54 90 57 20

e-mails: p.phelan@nature.com

k.turner@nature.com

US Head Office, New York

345 Park Avenue South,

10th Floor, New York, NY 10010-1707

Tel +1 800 989 7718

Fax +1 800 989 7103

e-mail: naturejobs@natureny.com

US Sales Director:

Ben Crowe

US Sales Manager:

Payton Mason

US Advertising Coordinator:

Ashly de Leon

Japan Head Office, Tokyo

MG Ichigaya Building (5F),

19-1 Harakatemachi,

Shinjuku-ku,

Tokyo 162-0841

Tel +81 3 3267 8751

Fax +81 3 3267 8746

e-mail: k.cowan@naturejpn.com

Japan Manager:

Kate Cowan

naturejobs

Cutting into the triangle

During a visit to North Carolina's Research Triangle — an area delineated by Duke University, the University of North Carolina at Chapel Hill, and the State University of North Carolina — conversations I'd had with professors and biotech employees gave me a specific impression. It seemed that scientists were being drawn to the area not just by the research activity, but also by the lower cost of living and easier commutes compared with other research-rich areas such as California and the 'Northeast corridor' (see pages 4–5). But I needed further evidence — perhaps a random sample.

Fate — and my clumsiness — conspired to give me just that. After leaving an interview at Duke University, I somehow managed to slice open a finger on a paper-towel dispenser. Fortunately, I was near to Duke's medical centre and, even more fortunately, a nurse there, sensing my perhaps misplaced urgency to make my next appointment, found a free physician. Paul Kuo, head of transplant surgery at the medical centre, even had the patience to tolerate a running tape recorder and my incessant questioning as he looked after the cut (I turned the machine off when he actually started stitching).

Kuo, it turns out, was recruited from Georgetown University in Washington DC about a year ago. He was attracted by the area's research climate — his next-door neighbour is involved in a biotech start-up — and Kuo no longer needs to justify time away from surgery, as Duke's hospital is not operated under a Health Maintenance Organization in the way that Georgetown's now is.

But Kuo notes that the intangibles are equally attractive. In North Carolina he has a larger house with more land than he could have afforded in Washington. And he doesn't have to load his childrens' bicycles into the car and drive them to a safe place to ride. "Life is so much easier here," he says.

Paul Smaglik
Naturejobs editor



Contents

REGIONS

North Carolina's
three-point plan p4

WWW.NATUREJOBS.COM

Career centre
Information on the
scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

SCIENTIFIC ANNOUNCEMENTS

SCIENTIFIC EVENTS



Building the triangle North Carolina

Construction time: facilities such as the North Carolina Biotechnology Center (above) and the University of North Carolina's new bioinformatics building (below) are helping to attract talent to the region.

The points of North Carolina's Research Triangle are formed by three universities — Duke in Durham, North Carolina State in Raleigh and the University of North Carolina (UNC) at Chapel Hill. But it is the interactions between these academic institutions and the local businesses and non-profit organizations that have driven the area's growth in terms of investment and jobs.

About 12 years ago, the Raleigh–Durham region of North Carolina had attracted only \$25 million in venture capital, and its 7,000-acre science park housed 32,000 employees. Now the triangle has attracted over \$1 billion in venture capital and more than 50,000 employees work within the park.

Recruitment and investment have increased because of the area's commitment to building new facilities and maintaining a high quality of life. It is taking an increasingly organized approach to attracting, nurturing and sustaining scientists in their entrepreneurial endeavours.

UNC's recruitment of William Snider from Washington University in St Louis provides a case study. The university first tempted him with plans for a new research lab, which his department moved into this year. Then the North Carolina Biotechnology Center — which has so far helped to bring 46 faculty members to the area's seven universities — sweetened his financial support package. Good opportunities for collaboration with nearby



A mirror to national trends

The uneasy financial climate slowed Research Triangle Park's growth last year. But the area's diversity has helped to keep it viable.

"If we were only an IT park, we'd be having a tough time," says Gary Shope, vice-president of marketing for the Research Triangle Foundation. The diversity keeps the park "insulated", he adds.

A look at employment trends among a few of the occupants may well reveal a microcosm of those in the United States generally.

GlaxoSmithKline, like

many other drug firms, is looking for medicinal chemists and animal pharmacologists. Syngenta has no plans to expand its scientific staff, even though its existing scientists will soon move into new labs and offices.

And Cisco Systems, which had plans to occupy more space in the park, is holding off until the economic climate improves. But the Research Triangle Institute, the park's fourth-largest tenant, has continued to expand — perhaps because it is a non-profit organization. **P.S.**

biotech companies such as Cogent in Durham, which specializes in therapeutics for neurological disorders, helped to close the deal. Those things, in turn, have helped Snider to fill up his new space. He has already recruited several younger staff members, almost all of whom are looking for postdocs and technicians.

The new building, which includes two floors of mice and roomy labs with windows, has been another draw. Three new hospitals in the past five years have



DUKE UNIV.

also helped others on the campus to attract scientists. Snider anticipates that a bioinformatics building and biological wet-lab building under construction will also prove attractive, as will a functional-genomics infrastructure being built for the park by IBM with investment by other local park inhabitants, including universities and the North Carolina Biotechnology Center.

The University of North Carolina isn't the only one enjoying a building boom. Duke is constructing a genetics building, an animal

facility and an engineering building — a third of which will be devoted to biomedical engineering. And North Carolina State University continues to expand its engineering facilities in order to attract high-tech companies such as Red Hat, the Linux software developer.

The quality of life in North Carolina has provided another magnet, Snider says. The newest addition to his faculty couldn't afford to buy a house in the Bay Area near the University of California, San Francisco, while he was doing his postdoc there. But he bought one just a short drive from the university soon after moving to North Carolina in January.

QUALITY ISSUES

Robert Taber, Duke University Medical Center's vice-chancellor for science and technology development, agrees that quality-of-life issues help the area to compete for top scientists. But he also wants to see a more entrepreneurial culture in North Carolina. He notes that potential recruits to Duke ask about business opportunities. They are often interested in seeing how the university can help them form their own company or create relationships with established organizations.



DUKE UNIV.

Taber's office can play a role in business development by helping faculty members to obtain patents, and then advise whether that intellectual property offers a licensing opportunity or the basis for a company. The area's Council

for Entrepreneurial Development also provides assistance. But the missing link, Taber says, is local venture capital.

Universities in the Bay Area of California can attract investment an order of magnitude larger than the entire Research Triangle area does. But the Californian investment is often locally based, meaning the venture capitalists are more likely to help the academics develop business plans and form management teams. It also means that they tend not to bail out so soon if the companies don't immediately

blossom. "If there's local venture capital, it makes all the difference," says Taber.

Another advantage North Carolina holds, especially in attracting larger companies, is its relatively inexpensive land, says John Reardon, who is GlaxoSmithKline's senior vice-president of discovery research in biology. Biotech pioneer Biogen, based in Cambridge, Massachusetts, has set up a branch in Research Triangle Park for that reason.

The relatively low price comes with some caveats. Research Triangle Foundation, which owns and operates the park, requires occupants to build on only about a third of the acreage they occupy and to keep buildings a certain distance from all roads. These rules create a green atmosphere for employees. But such stipulations make it difficult for smaller companies to move in. Nevertheless, many start-ups are managing to find space, either with help from non-profit organizations, or by seeking land outside the park itself.

AlphaVax — which, in conjunction with the US National Institutes of Health, will have an RNA-derived AIDS vaccine in clinical trials this spring — was initially located in Durham, after having its roots in research at UNC Chapel Hill in 1989. But a loan from the North Carolina Biotechnology Center helped the company to move to the park.

Trimeris, another local biotech company, has benefited from being in the area. The company was born in the lab of Dani Bolognesi, an AIDS researcher at Duke. The university helped him spin out the company and gave him leave of absence to develop it. In terms of recruitment, AlphaVax and Trimeris are in the same boat. They are not likely to take on more scientists until they hear positive news about their clinical trials.

Taber notes that the success of such small biotech companies will determine whether the area evolves from simply hosting mature companies to serving as a base for the growth of new ones. If companies spun off from Duke, UNC Chapel Hill and North Carolina State universities are successful, venture capital will flow into the area more freely, and job growth will follow.

Paul Smaglik is editor of *Naturejobs*.

North Carolina Biotechnology Center ♦ www.ncbiotech.org

Research Triangle Foundation ♦ www.rtp.org

Council for Entrepreneurial Development ♦ www.cednc.org

Expanding: Duke University is investing in new buildings.

Robert Taber: wants to encourage an entrepreneurial approach in North Carolina.

Signing a contract

Some of the world's largest contract research organizations call Research Triangle Park home. Companies such as Quintiles and Covance offer jobs to PhDs who don't find bench work exciting and MDs who want to manage somebody else's clinical trials.

But now, this phenomenon is showing signs of extending beyond clinical

medicine. The local branch of Dutch company Diosynth, which helps small biotechs without high production capacity to develop and manufacture proteins for drugs, has grown from six staff to 600 in six years. And Scynexis, a younger company, offers support in medicinal chemistry. Both companies are seeking new employees.

P.S.